Analyzing the Distribution of Energy Sources in the United States

Suvrath Arvind, Clayton Greenberg

Abstract

The United States is one of the largest consumers of energy in the world, but this energy comes from a wide variety of sources. In addition, this energy consumption varies from state-to-state and from sector-to-sector, meaning that no one model would tell us the full story of the energy distribution in the United States. The goal of this project was to analyze this data, using various techniques to develop our understanding of the nature of the data we were provided with. To effectively analyze the data, we created three groups of data: individual energy sources, energy by state, and general aspects of the energy distribution. Our analysis showed us that energy sources, like coal, appeared to be decreasing in consumption, that states could be grouped in clusters in order to predict production of coal from consumption (coal was the main energy source we analyzed), and that other aspects of the energy distribution (consumption and expenditure, for example) were almost perfectly correlated.

1 Introduction

The use of nonrenewable energy sources, especially coal, in the United States has proven to be a major issue. It can be assumed from common knowledge that burning fossil fuels is harmful to the environment, and decreasing both the production and consumption of those energy sources is essential to keep the world thriving naturally. Our analysis of the energy sources showed that even today, the use of nonrenewable energy sources is still far more common and that the use of these sources does not seem to be going away anytime soon. While traditional nonrenewable energy sources are decreasing nationally, certain states are still major producers and consumers of these. Legislation has been passed at both the state[1] and federal[2] levels in an attempt to curb the production and consumption of these energy sources, but the effects would only be seen in decades.

Predictive analysis in energy distribution is extremely important as this is the main way to see how the country is doing to curb the ongoing nonrenewable energy crisis. However, given that energy is a broad topic, we had to figure out what aspects of the energy distribution can be grouped together in order to find as many robust and distinct patterns as possible. When it came to predicting the production of certain energy sources (in this case, coal), we could use a least squares regression approach that would involve polynomials. This approach would be enough given that predictions could be made about the near future with decent accuracy and would not be too sophisticated to overcomplicate the problem. Using time as a factor in the predictions, we

[1] https://www.energy.ca.gov/rules-and-regulations/energy-suppliers-reporting/clean-energy-and-pollution-reduction-act-sb-350
[2] https://afdc.energy.gov/laws/key_legislation

would be able to approximate the amount of coal produced in a certain year and then use that to further our analysis.

However, using predictions, by year, would not give us helpful information on other aspects of the energy distribution. Instead using a heatmap, a feature that would display the correlation between all aspects of the energy distribution, would give us a better picture of how individual aspects were all related to one another. When developing a heatmap, we used two approaches: using Pearson Correlation and using Spearman Correlation. Both of these approaches are similar to regular correlation, but each has unique aspects. For instance Pearson Correlation would give us information regarding the correlation between two variables, by looking at each value equally, whereas Spearman Correlation would compare variables by rank. Rank, in this situation, would be the position of a certain data point, meaning that if there were two data points at the same position, they would have the same rank. This was helpful, given that there were some states that shared certain data points. In short, Pearson requires more data points, but it can give a more significant result quicker.

Another approach that we took to accurately analyze the energy distribution was using clustering. This approach essentially groups closely related data points together and uses that to organize the data. This was helpful since we could then make a map of the United States and color in the states based on the cluster they fell into which would give us a visual representation of where certain aspects of the energy distribution were higher and how geography could have some role in predictions. We could also use the clusters to make predicting consumption from production easier by essentially splitting the problem down by the number of clusters we had.

2 Background

Clustering, in general, is a machine learning technique that groups data points based on where they are located when plotted on a graph. For analyzing the coal distribution, we used a K-Means clustering technique, a technique which uses *k*, the target number of clusters (or groups) and finds *k* centers (or means) of these groups.[3] The number of clusters are chosen by the user rather than the machine. The centers should represent as many data points as possible and the K-Means approach does so by making circular clusters. This is the simplest clustering technique since this method is told what to find and it finds it for us. However, this technique is limited in that it does not perform that well when it comes to data points that form lines or rays as linear data points cannot be grouped in circular clusters.

When using a polynomial model to predict the coal distribution, we used Polyfit, a feature of the NumPy library of Python. This uses a least-squares regression model to represent the data and is given the degree of the polynomial used to predict the data. This model is quite accurate up to a fifth-degree polynomial, and after that the model gets too complex and results in a very inaccurate model. While this may be counterintuitive, higher degree polynomials should be able to capture more data points, the Polyfit feature is limited in its abilities when

[3] https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

forced to produce a polynomial model with a very large degree and generates results worse than even linear model when used to represent the data, which is quite complex. However, when producing models with degree 3, 4, or 5, the model was relatively accurate.

Lastly, we used coal as our primary energy source we analyzed since coal has been historically used the most to generate power in the United States. Analyzing coal would give us a better picture of the distribution of energy sources in the United States as changes in the coal distribution would result in changes in all the other distribution of energy sources.

3 Dataset

The data used for this project came from Corgis's Energy Python Library[4] and it provided us with the energy distribution of each state (and Washington D.C.) in the United States per year. The energy distribution itself was made up of 84 different groups, with each group categorized by the type of energy, where that energy was being used, and what aspect of that energy was being studied. For example, the group representing the consumption of coal in the residential sector would be represented by the group labeled "Consumption.Residential.Coal."

The dataset included 3060 energy distributions, distributions for each state and Washington D.C. from 1960 to 2019, meaning that only using a part of the dataset would be the only way that analysis could be done. To take care of this situation, we made groups that combined the energy distribution for all the states for each year and only considered some of the years. This was helpful in that we could consider the energy distribution for every year without worrying about the states' individual distributions. To base our predictions, we used the 2019 energy distributions, as this was the last year recorded in the data set and would be the most accurate in predicting future values. One advantage of having the year that we analyze as a constant was that we could see how each of the 84 aspects of the energy distribution were related, connections that might have not been obvious from the outside.

We also made groups that represented the energy distribution for each energy source. This allowed us to see how distributions for specific energy sources changed over time, by considering the United States as one whole entity, without the individual contributions to the energy distribution by the states as a factor. In this step, we focused our attention onto coal as this was the only one that had data points regarding production, meaning that using this could allow us to see how production and consumption, two aspects of the energy distribution that has greatest effects on the United States as a whole, were related to one another. Grouping by consumption and production was the basis of the clustering method and this allowed the data to be transformed from having 84 different groups to only 2.

We also focused on petroleum in some of the analyses to see how reliable some models were in predicting different parts of the energy distribution, and we also used this energy source to improve our understanding of the different types of energy sources present in the dataset, rather than solely focusing on one, but even then, coal was analyzed the most due to the

[4] https://corgis-edu.github.io/corgis/python/energy/

presence of a “Production.Coal” category as the coal distribution provided the most information. Below are the graphs representing how the production and consumption of coal changed over time, and how the consumption of petroleum changed over time. The latter was used in the first aspect of our analysis: using polynomial models for prediction.

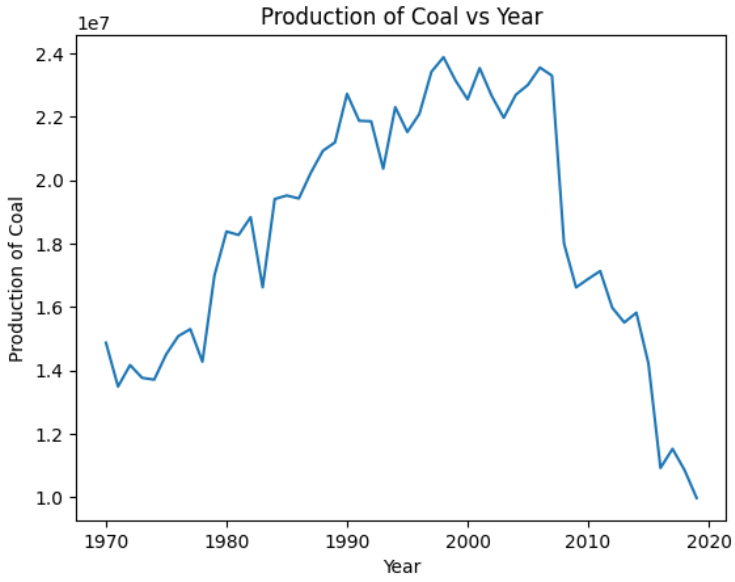

Fig 1: Graph showing how production of coal has changed over time

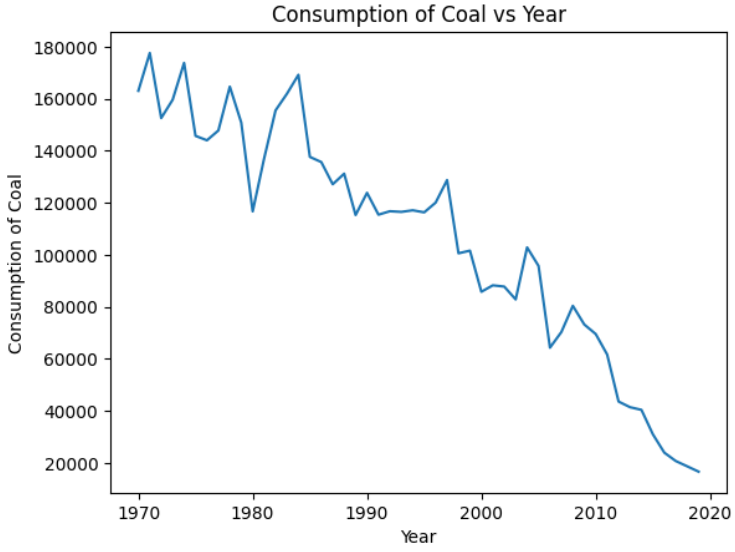

Fig 2: Graph showing how consumption of coal has changed over time

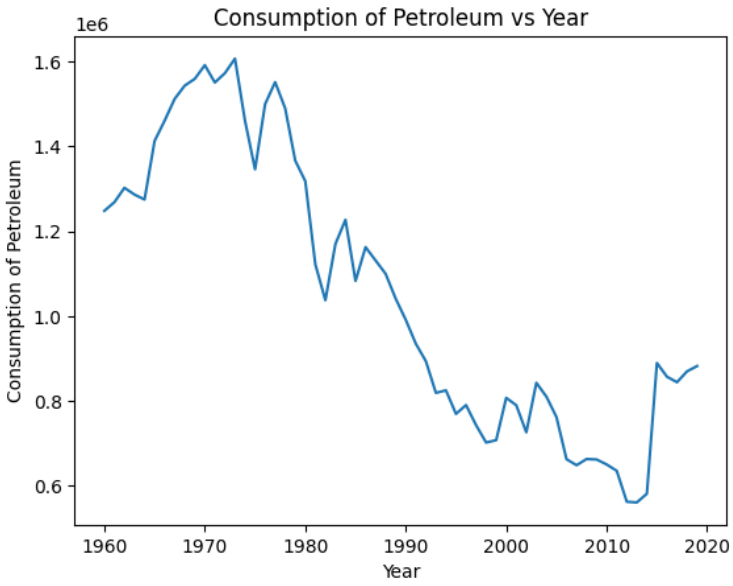

Fig 3: Graph showing how consumption of petroleum has changed over time

By focusing on coal (and in some cases petroleum) and on the 2019 distribution, we were able to get the data transformed enough to make analyzing the data as straightforward as possible and it limited any possibility of error in the analysis being conducted.

4 Analyzing the Dataset

4.1 Energy Sources Over Time

When it came to predicting energy sources over time, we used polynomials to model the data. Instead of focusing on every energy distribution, we focused on seeing how the consumption of petroleum changed over time (**Fig 3**) in order to predict future values of petroleum consumption. The accuracy of the model was determined by using the p-value determined by Shapiro-Wilk Test. This test checks for Normality in the distribution of residuals, as a normally distributed residual plot would indicate that the model was a good fit for the data. Unlike most hypothesis tests, a larger p-value (indicating that there is no convincing evidence that the alternate hypothesis is true) is desirable as the null hypothesis for this test is that the residual distribution is in fact normally distributed.

To create these polynomial models, we used Polyfit, a method supplied by Python that uses least-squares regression to develop the desired model. For each model, we determined the Shapiro Test p-value (to determine the validity of a model), and assessed the quality of each using the Mean Absolute Error (MAE). We calculated the MAE for the polynomials that resulted in a high Shapiro Test p-value and used that as a metric to confirm that a certain model was the best. We started with a first-degree polynomial and increased the degree of the regression curve until we found invalid results. The table below summarizes the results.

Degree	p-value	MAE
1	0.0072	N/A
2	0.068	136216
3	0.93	67886
4	0.14	62245
5	0.56	61687
6	4.5×10^{-12}	N/A
7	1.2×10^{-10}	N/A

Table 1: Table showing the p-values and MAE for each model based on degree

As the table shows, polynomials of degree 4 and degree 5 performed the best in predicting the consumption of petroleum due to their low MAE values. This can be seen in the following graphs where the predicted curve closely matches the original curve.

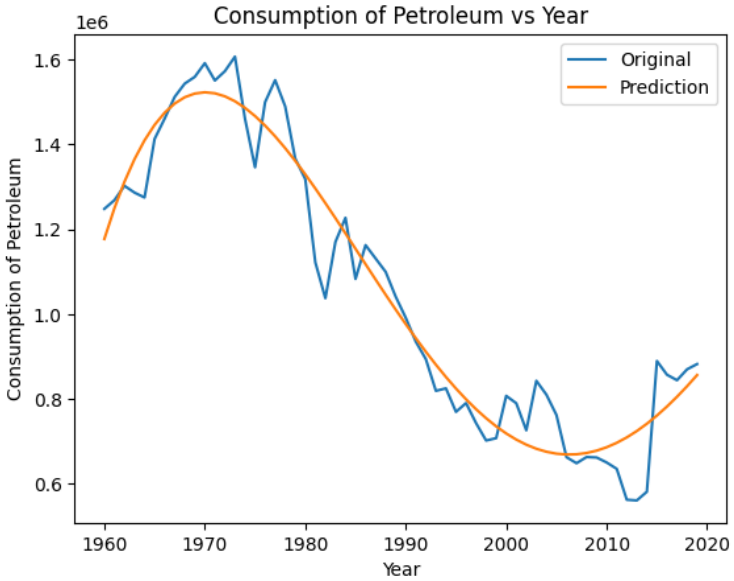

Fig 4: 4th Degree Polynomial used to predict consumption of petroleum

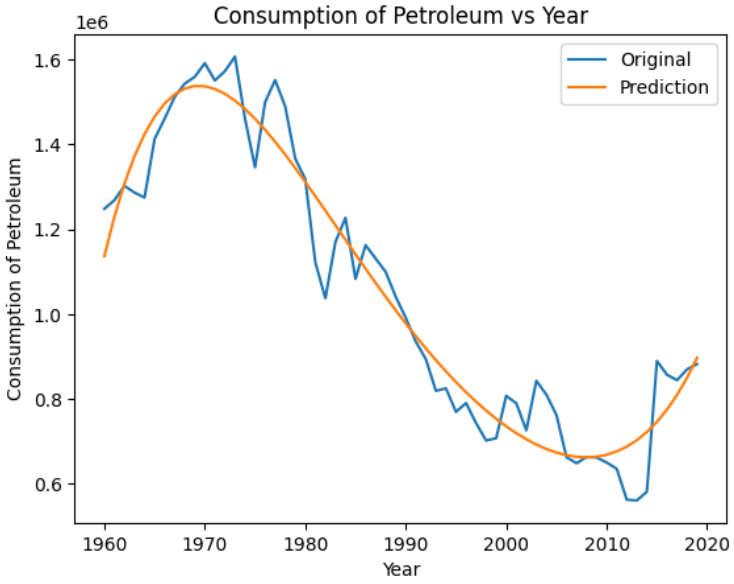

Fig 5: 5th Degree Polynomial used to predict consumption of petroleum

As the two graphs show, the fourth-degree and fifth-degree polynomials performed the best, and while it may see that the sixth-degree polynomial should perform better, that is not the case due to the lack of validity provided by the Shapiro Test. Since the p-value of the Shapiro Test did not indicate which model was the best and only the MAE did, the fourth-degree polynomial best represented the data, with the fifth-degree and third-degree polynomials following.

4.2 Aspects of the Energy Distribution

4.2.1 Finding Outliers Using Maps

In order to further understand the data, especially in order to find outliers, we used a map that would indicate to us what state produced or consumed the most amount of a certain energy source. Since production values were only given to coal, we used the coal distribution when creating the maps. These maps were just maps of the United States, but each state was colored differently based on the amount of coal produced or consumed. For example large producers or large consumers were labeled in a more yellow color, whereas small producers or small consumers were labeled in a more purple or blue color. This can be seen in the following diagrams, which represent the production of coal and consumption of coal, respectively.

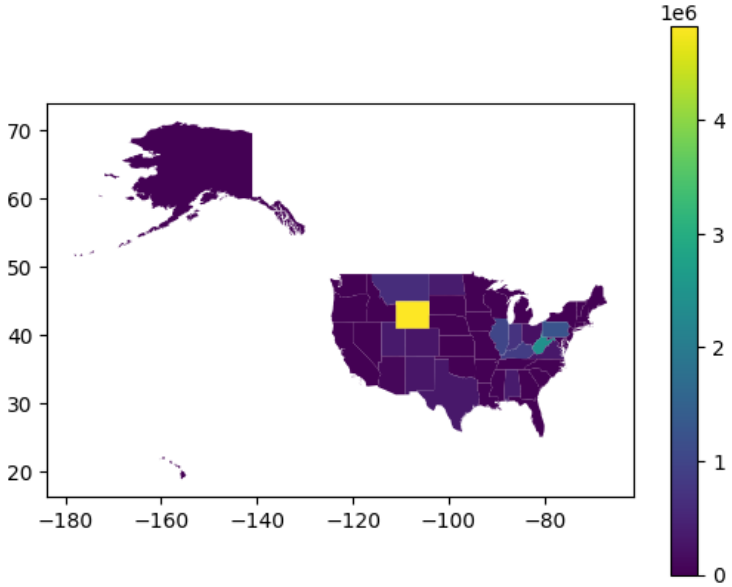

Fig 6: Map representing coal production in the United States

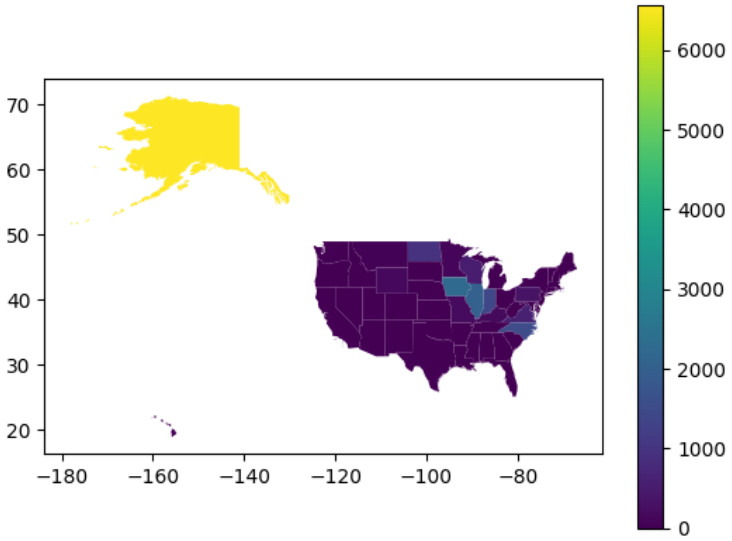

Fig 7: Map representing coal consumption in the United States

As the data show, Wyoming is an obvious outlier in the coal production aspect of the dataset, while Alaska is an outlier in the coal consumption aspect of the dataset

4.2.2 Analyzing The Data Using Heatmaps

The next approach we took to analyze the energy distribution was to use heatmaps to find if there was any correlation between various aspects of the energy distribution. In order to make our analysis easier, we looked at the energy distribution for 2019 as keeping the year constant allowed us to see if there was any relationship between variables of a certain year.

There were two heatmaps that we used: one that used a Pearson correlation and one that used a Spearman correlation. The difference between the two was that Pearson considered each data point equally (not taking into account the location of the data point) and Spearman considered the rank of each data point (points at the same location would have the same rank). Below are the heatmaps for Pearson and Spearman, respectively.

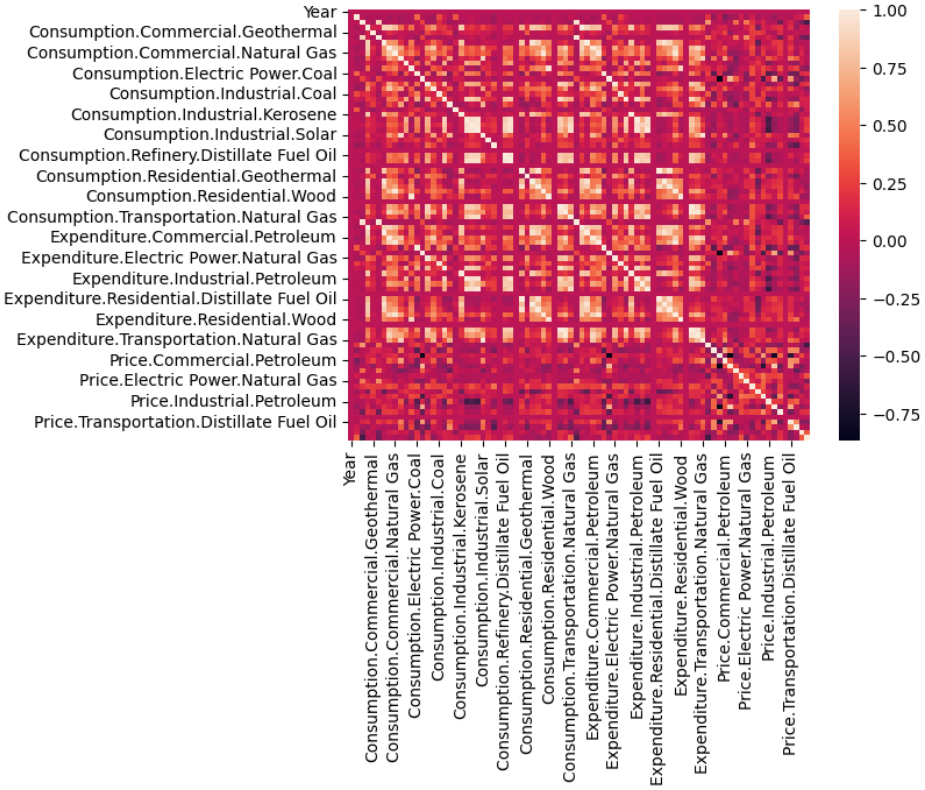

Fig 8: Heatmap produced by Pearson Correlation

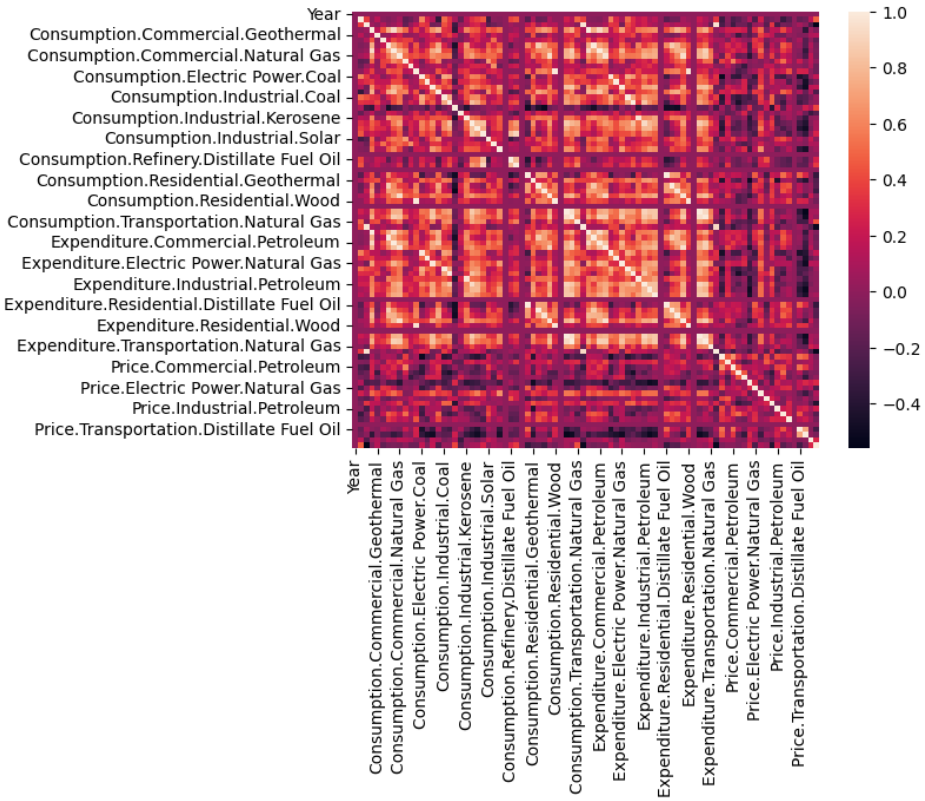

Fig 9: Heatmap produced by Spearman Correlation

As the two heatmaps show, there is a strong correlation between various variables as seen by the very dark squares and very light squares in both distributions. The difference between the two heatmaps is caused by the implementation of each of the correlation methods. The Pearson heatmap looked like a more uniform color, with squares (which represented the correlation between the two variables each square represented) in a more muted color, indicating that correlation was closer to 0. The Spearman heatmap, on the other hand, produced more distinct squares as repeated points were considered as one, meaning that there would be fewer points to base the correlation on as many of the points were at (0,0).

What stood out from the heatmaps was that there appeared to be scratches in the heatmap, places where there appeared to be lines representing areas of high correlation. The middle “scratch” was the only one that was expected as any variable is perfectly correlated with itself, while the “scratches” seen on the outside of the heatmap were surprising. This led us to a hypothesis that there is some relationship between variables that might not seem connected from the outside.

One obvious “scratch” in the heatmap was that squares that represented Consumption and squares that represented Expenditure had a strong correlation. When plotted, that correlation becomes obvious.

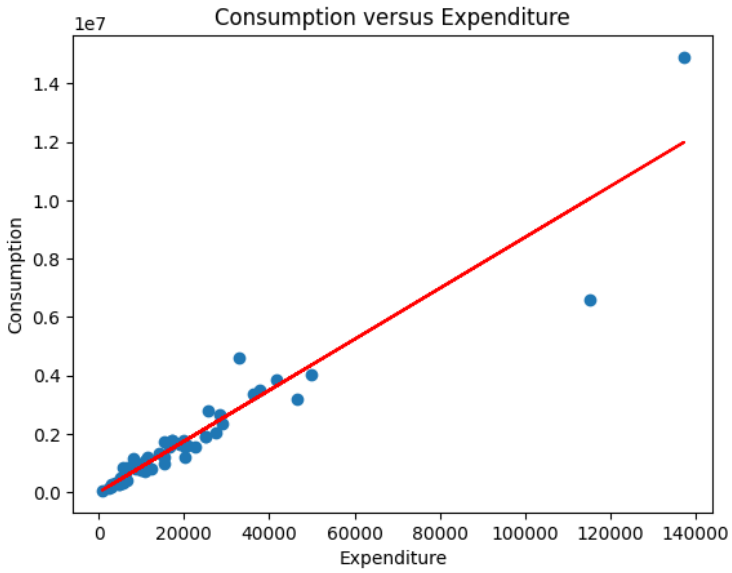

Fig 10: Predicting Consumption from Expenditure

As this graph shows, there is a relatively strong correlation between Consumption and Expenditure in the 2019 energy distribution. This is seen by the line of best fit closely matching the data points on the left side of the graph, with only a few outliers not being able to be represented by this line. The two obvious outliers represent California and Texas, the two most populated states in the United States, so these two states being outliers is expected.

In addition, some of the scratches were more obvious. For example, the heatmap shows that there is a strong negative correlation between Price and Consumption. This makes sense since increasing the price of certain energy sources would end up decreasing the consumption of that energy source, and increasing the consumption of a specific energy source would decrease that energy source's price.

Strong correlation between other aspects of the energy distribution allowed us to understand that just looking at certain aspects, like production and consumption, does not give us the full picture of the entire energy distribution, but instead looking at aspects like expenditure and price can unearth some interesting features of this dataset.

4.2.3 Comparing Renewable Energy Sources with Nonrenewable Energy Sources

One more aspect of the energy distribution that we analyzed was the consumption of renewable and nonrenewable energy sources. Renewable energy is a more sustainable form of energy, and seeing if the United States is on a trend to use more renewable sources than nonrenewable sources is important. Using the 84 groups given in the original dataset, we created subgroups of data for each energy source and then combined groups that represented renewable energy sources and groups that represented nonrenewable energy sources. Renewable energy sources included geothermal, hydropower, solar, and wind; nonrenewable energy sources included coal, distillate fuel oil, kerosene, petroleum, natural gas, and wood. By

combining these sources, we were able to plot both the consumption of renewable energy sources and consumption of nonrenewable energy sources over time together, as seen in the figure below.

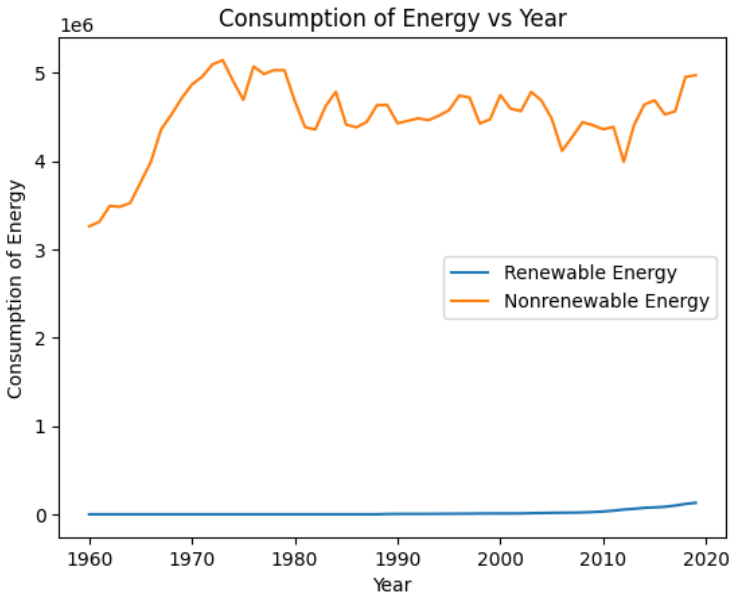

Fig 11: Graph showing the consumption of both renewable and nonrenewable energy sources over time

As the graph shows, the consumption of renewable energy sources looked almost negligible to the consumption of nonrenewable energy sources. This was expected as many Americans still consume nonrenewable energy sources, but this staggering difference is still quite shocking. In addition, this graph shows us how renewable energy sources only started to be consumed in the early 2000s, meaning that there is no chance that renewable energy sources would dominate nonrenewable energy sources when it comes to consumption in the near future.

4.2.4 Grouping By Sector

Another aspect of the energy distribution that we analyzed was how each sector uses energy. We did this by making a bar chart, as seen in the figure below, representing how much energy was consumed in each of the sectors of the energy distribution.

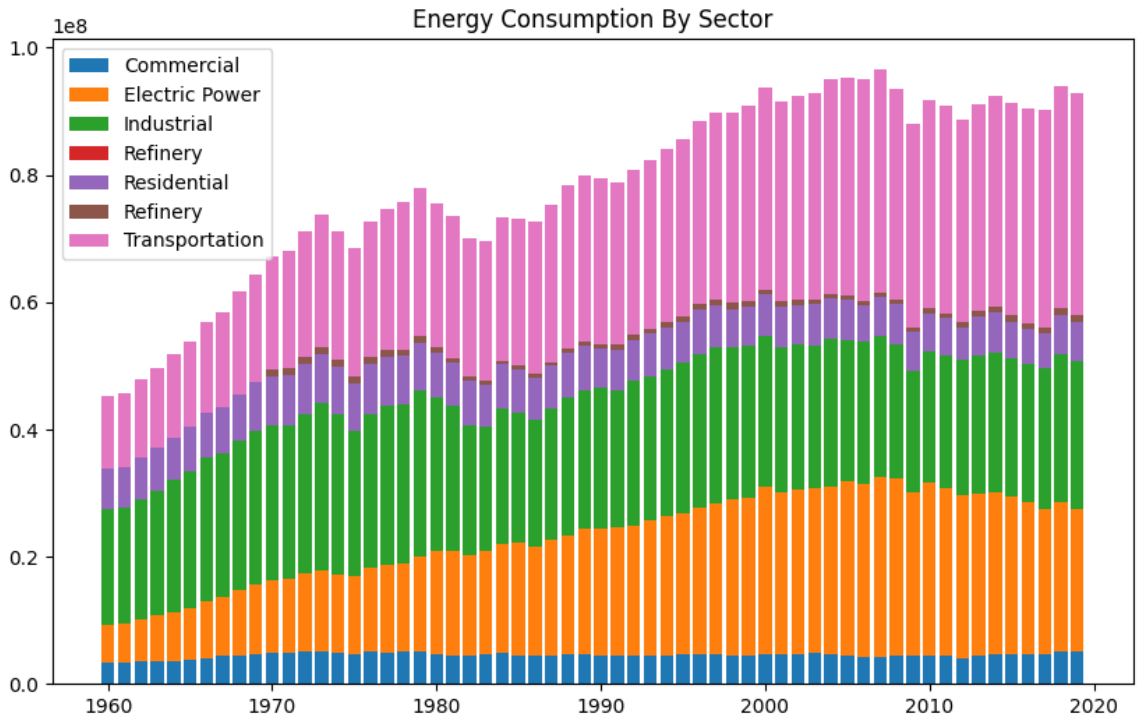

Fig 12: Bar chart showing energy consumption by sector

This bar chart shows how certain sectors use energy more than others and how energy consumption, in general, has been growing steadily. This also shows that in some sectors, like Transportation, the consumption of energy has been growing, unlike the Commercial sector, where energy consumption has been relatively stagnant. This bar chart is a good visualization of how energy consumption has been changing over time through the lens of various sectors, an aspect of the energy distribution that was not heavily analyzed.

4.3 Predicting Consumption from Production Using Clustering

4.3.1 Developing the Clusters

The last technique we used to do predictive analysis on the energy dataset was to use a clustering since, when plotted together, consumption and production of coal seemed to form groups, as seen in the figure below.

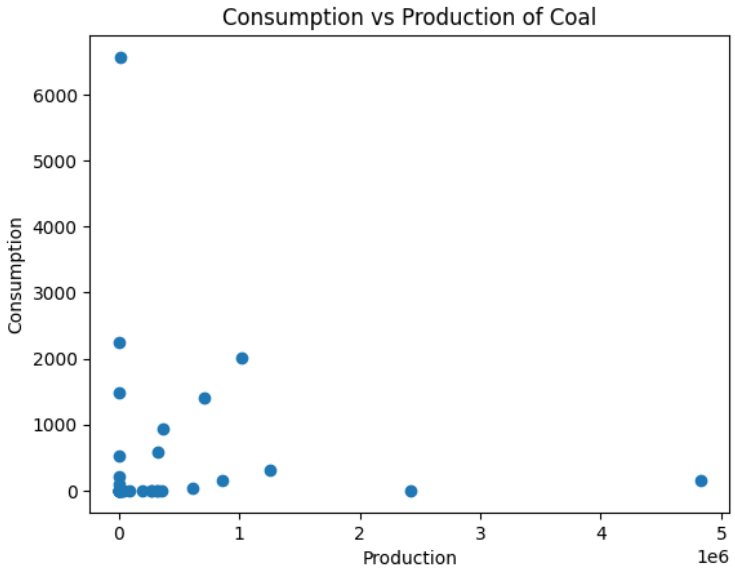

Fig 13: Graph of Consumption vs Production of Coal

As the graph shows, there are distinct groups in the dataset forming what appears to be rays from the origin. As explained earlier, a fixed number of groups needed to be determined to allow for K-Means clustering to take place, so based on this graph (**Fig 12**), we decided that 4 clusters (or groups) would be necessary. However, this did not give us favorable results, as seen in the following figure, so we were forced to use manual clustering to actually get the clusters to be relatively accurate predictors.

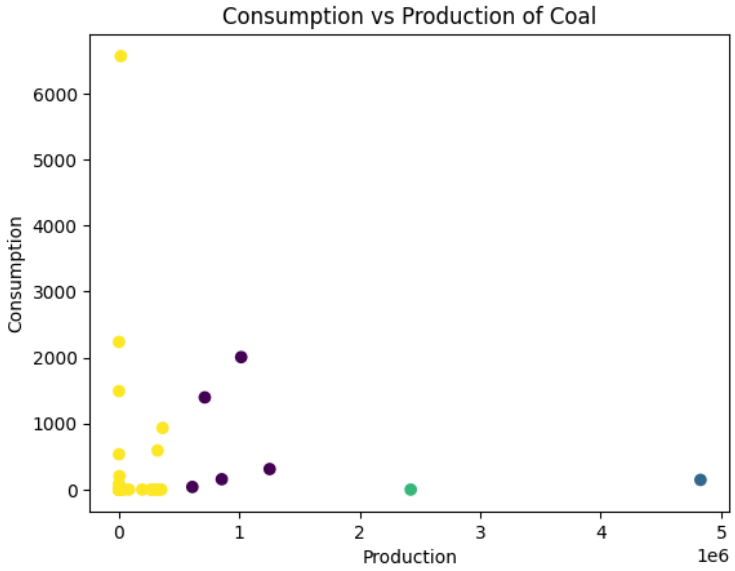

Fig 14: Initial clustering output

As the graph shows, this clustering did not capture the data points in the way that we wanted to, which was expected. The model found 4 centers, as it was told to do, and chose the

centers in a way that each center was as close to as many data points as possible. This meant that using the K-Means clustering technique was not the most accurate. After seeing this result, we decided that manual clustering, which assigned each data point a value, was the way to go, and we created 7 clusters (from Cluster 0 to Cluster 6). Cluster 0 (represented a state that produced very little coal and consumed a lot of coal, while Cluster 6 represented states that produced a lot of coal, but consumed very little. The clusters in between would represent states that were in the middle. The figure below represents the new scatter plot with clusters that were manually created.

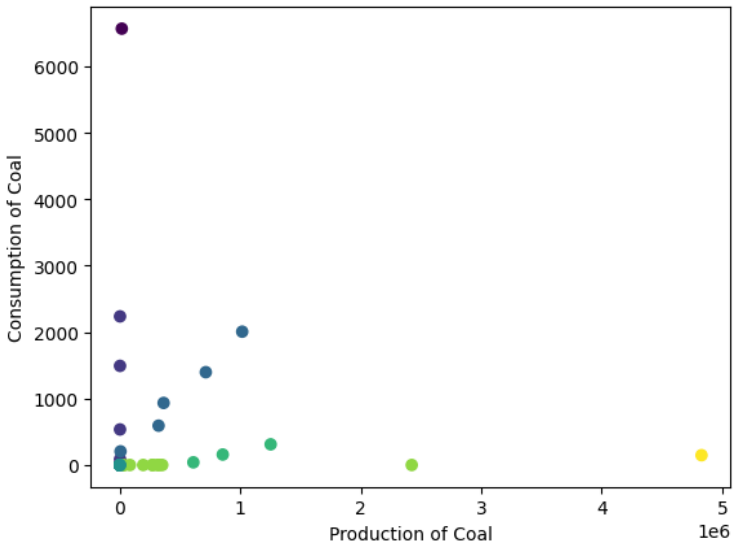

Fig 15: Manual clustering output

As the scatter plot shows, categories generated by this approach were a better representation of the data. Cluster 0 is represented by the dark blue dot, while Cluster 6 is represented by the yellow dot. This is further seen in a map representation of the scatter plot.

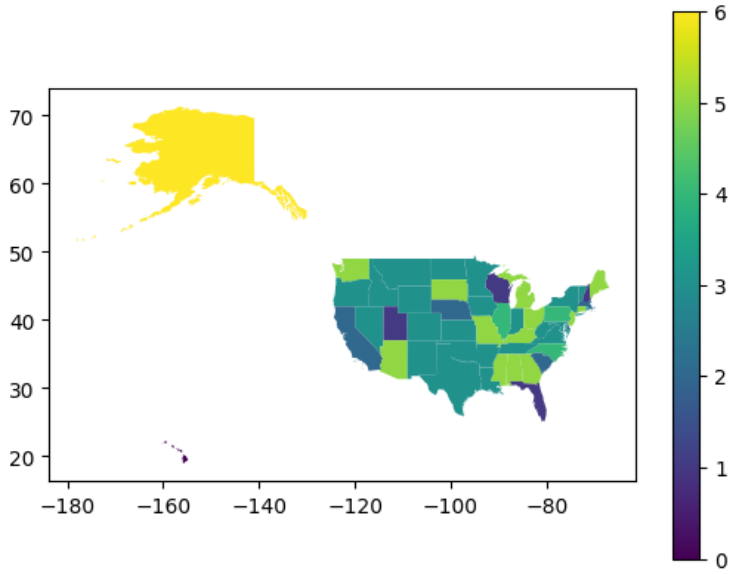

Fig 16: Map representing the clusters

The map is essentially a combination of the two individual maps that represent the consumption and production of coal per state in 2019. The manual clusters showed how each state did when it came to producing and consuming coal and indicated that geography could also be a factor when it came to producing and consuming coal.

4.3.2 Using Clusters To Predict Consumption From Production

The clusters that were created were then used to predict coal consumption from production and various combinations were used:

Model 1: Predicting future consumption (represented by Newc) from past consumption (represented by Oldc)

Model 2: predicting future consumption from past consumption added with current production (represented by Newp),

Model 3: Predicting future consumption from past consumption added with the cluster number (represented by C(Cluster)) multiplied with the current production.

The models were produced by using OLS (Ordinary Least Squares) regression. The AIC of each model was calculated, and a lower AIC value would indicate a better fit. A 95% confidence interval was also calculated and the p-value was used to determine significance. Below are the results for each of the three models described previously.

AIC	590.9			
Explanatory Variable	Left Endpoint	Middle	Right Endpoint	p-value

Oldc	0.90	0.92	0.94	0.00

Table 2: Table showing the results of Model 1

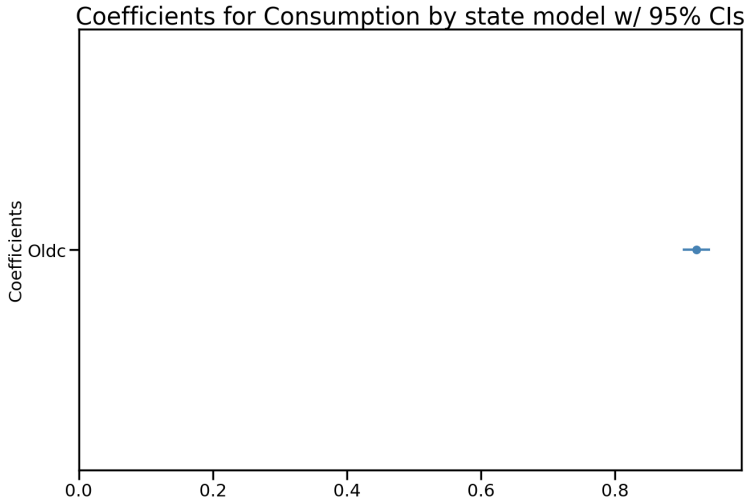

Fig 17: Confidence Interval created for Model 1

AIC	592.5			
Explanatory Variable	Left Endpoint	Middle	Right Endpoint	p-value
Oldc	0.90	0.92	0.94	0.00
Newp	-3.8×10^{-5}	-9.5×10^{-6}	1.9×10^{-5}	0.51

Table 3: Table showing the results of Model 2

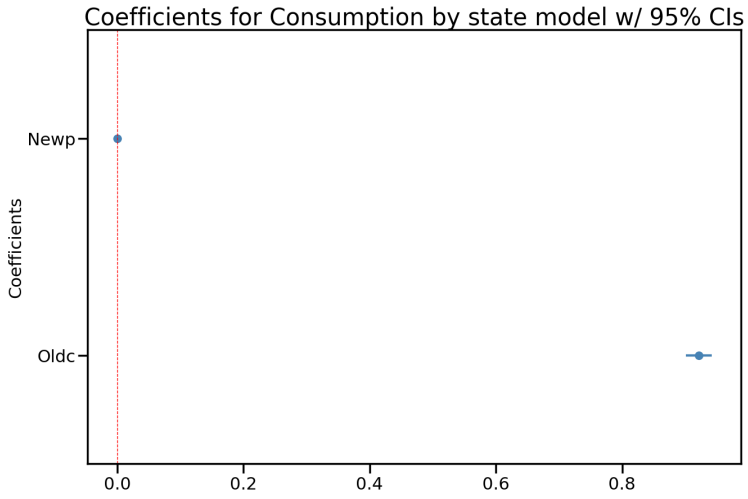

Fig 18: Confidence Intervals created for Model 2

AIC	577.4			
Explanatory Variable	Left Endpoint	Middle	Right Endpoint	p-value
C(Cluster)[T.1]	-80.82	2.54	55.53	0.95
C(Cluster)[T.2]	-59.87	33.53	85.89	0.47
C(Cluster)[T.3]	-60.14	-6.32	126.94	0.81
C(Cluster)[T.4]	-238.36	-17.10	47.49	0.88
C(Cluster)[T.5]	-65.52	-6.32	204.15	0.83
C(Cluster)[T.6]	-1.6×10^{-8}	-9.8×10^{-9}	-4.2×10^{-9}	0.001
Oldc	0.79	0.84	0.90	0.00
Newp	0.02	0.05	0.08	0.001
C(Cluster)[T.1]:Newp	-2.1×10^{-14}	-1.4×10^{-16}	-2.08 × 10-14	0.99
C(Cluster)[T.2]:Newp	-0.08	-0.05	-0.02	0.001
C(Cluster)[T.3]:Newp	0	0	0	nan
C(Cluster)[T.4]:Newp	-0.08	-0.05	-0.02	0.001
C(Cluster)[T.5]:Newp	-0.08	-0.05	-0.02	0.001
C(Cluster)[T.6]:Newp	-0.08	-0.05	-0.02	0.001

Table 4: Table showing the results of Model 3

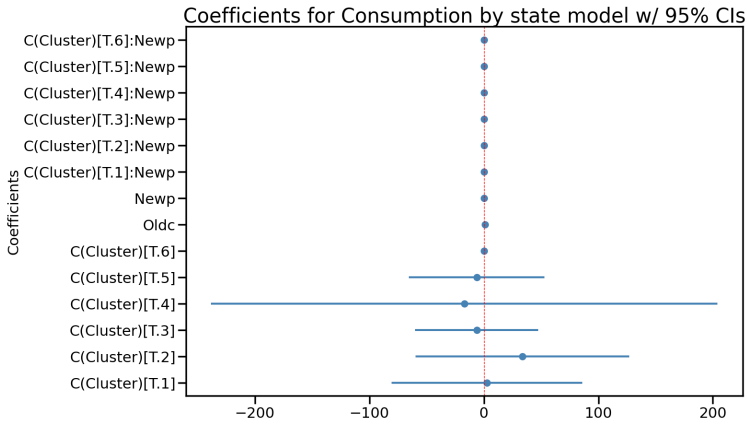

Fig 19: Confidence Interval created for Model 3

As the figures and tables show, the best model was Model 3 (the only one that used clusters) due to its AIC (577.4) being the lowest out of the three models. There were some significant results especially in the terms that involved an interaction (represented by ":" in the table). While it might seem like some of the confidence intervals for the interaction terms crossed 0 in the graph (**Fig 20**), that is not the case when looking at the table (**Table 4**), meaning that these terms in fact could be used as the best predictors as every other confidence interval had 0 as a plausible value. The reason why 0 should not be captured is that a coefficient of 0 (the confidence intervals are representing plausible coefficient values) would mean that there is no correlation, resulting in a bad fit.

5 Conclusion

In conclusion, this project showed us the complexity of the distribution of energy sources in the United States. Energy is a broad topic, with various aspects to it, and in this project, we were able to touch on as many aspects as possible. Analyzing this distribution showed us that each state was very different when it came to energy needs, but it also showed us that neighboring states can have similar distributions.

We were successful in using various approaches, from polynomials to clustering, to have a better understanding of the entire energy distribution and this allowed us to look into the distribution through a completely different lens: one that approached the data points first and then what the data points stood for. This enabled us to effectively analyze the data and allowed us to draw several conclusions about each of the aspects of the distribution of energy sources in the United States.

In the future, this analysis could be expanded even further to look for more sophisticated relationships between variables or this approach could be used in analyzing other aspects of the globe as a whole.

Acknowledgements

We would like to thank Polygence for enabling this project to happen. This platform allowed us to effectively discuss the dataset thoroughly, giving us a more complete picture in our analysis.

References

[1] *Alternative Fuels Data Center: Key Federal Legislation*. Alternative Fuels Data Center. (n.d). Retrieved May 13, 2023, from https://afdc.energy.gov/laws/key_legislation

[2] *Clean Energy and Pollution Reduction Act - SB 350*. California Energy Commission. (n.d). Retrieved May 13, 2023, from https://www.energy.ca.gov/rules-and-regulations/energy-suppliers-reporting/clean-energy-and-pollution-reduction-act-sb-350

[3] Education Ecosystem (2018, September 12). *Understanding K-means Clustering in Machine Learning*. Towards Data Science. Retrieved May 13, 2023, from https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

[4] Bart, A.C., Choi, J.M., & Guan, B. (2021, October 7). *Energy Python Library*. CORGIS Dataset Project. Retrieved May 13, 2023, from https://corgis-edu.github.io/corgis/python/energy/