Data 102 Final Project

Team 51

Gabriella Rosal Calit, Amitesh Gargapati, Mahathi Kattamuri, Ashwin Iyer

1. Data Overview

The data we used for our research comes from a dataset containing information on monthly electricity consumption across all 50 states over a 12-year time period (US Energy Information Administration). Data were drawn through a sample, and the dataset also includes information on variables such as customer, count, and prices. There were no missing values; no records were dropped. In conjunction with electricity consumption data, we also used data on several variables specific to New Jersey, our state of choice, and sourced economic data (CPI, unemployment rates) from the St. Louis FRED website by searching for our variable of interest on their database and adjusting the locality and temporality of the graph before downloading CSV files. Furthermore, we sourced weather data from NOAA (National Centers for Environmental Information) and Climate Data Online (CDO) in order to obtain information on variables such as temperature, wind speed, precipitation, etc. Information on these variables was not present in our initially chosen dataset, so sourcing it from other datasets was crucial in order to fully explore the relationships we wanted to study. Electricity consumption data was collected directly at the source; we did not see any need to impose differential privacy paradigms onto our analysis. We then combined this information based on the date (monthly observations) with our electricity dataset to create the dataset upon which we performed the entirety of our analysis. One of the variables sampled in our dataset was temperature, drawn across the 12-year period, and when compared to historical averages, the average temperature in New Jersey in our dataset (53.6 Fahrenheit) closely matches the US average temperature in roughly that time period (54.4 degrees Fahrenheit) and our choice of thresholds and our results will likely have some reproducibility when looking at the rest of the country. One thing to note is that New Jersey's temperatures tend to have more fluctuations than the rest of the country (more mass in the tails for a temperature distribution) so this may influence exactly how valid our results are for states that are not geographically situated in the same area as New Jersey.

Each row in our aggregated data set represents one observation taken at the start of the month, and reports that date with the associated electricity consumption (megawatt hours), average temperature, average wind speed, etc. We do not think selection bias or convenience sampling played significant roles in the data collection process, as every locality (NJ area) is equally represented and the entire state is covered. As for measurement error, our data averages for each variable look to be within the historical averages in each area, and while we do not have perfect domain knowledge to make claims about the specifics of how these data were measured, we can conclude for the purposes of our analysis that measurement error does not pose a serious concern.

Data that we would have liked to see as part of the dataset would have included information about the size of a household from which consumption data was collected, surge pricing information during peak summer or winter, or government transfers (like unemployment insurance) during this time period. This information would have helped us eliminate even more potential sources of omitted variable bias in our analysis.

Once we assembled our aggregate data, we standardized the date-and-time formatting in order to match our data into one dataset, and we set the key response variable (megawatt hours) to a numeric data type. Data cleaning will not impact our analysis and our conclusions. We then split our data into various thresholds for the subsequent propensity score analysis.

2. Research Questions

The first research question we want to investigate is **how weather affects electricity consumption**. Answering this question can inform several real-world decisions, especially when it comes to emergency preparedness. Government agencies and emergency first-responders can use weather-electricity consumption data to prepare for weather-related emergencies. In serious situations, having backup power is a necessity for essential services and critical infrastructure. In February 2021, the state of Texas suffered a major power crisis due to a series of snowstorms, resulting in hundreds of deaths. Our aim in answering this question is to understand if weather-electricity consumption data can be used to prevent similar catastrophes in the future. Utility companies could also use the data to enhance their demand response programs. These programs are meant to temporarily reduce electricity consumption when demand is greater than supply, usually during hot and humid days. Utility providers can use the weather-electricity consumption data to predict high-consumption days and offer incentives to reduce electricity use.

We will answer the first question using causal inference, particularly through propensity scores. The propensity scoring method helps address confounding variables such as seasonality. We will use this method for each treatment variable while re-weighting based on the other variables that are not being used as treatment. We will start by calculating propensity scores for each observation, and then we will group our observations based on their scores in order to compare electricity consumption across each treatment variable and determine which treatment variable has the greatest causal effect. We are confident that this method reduces bias and improves the accuracy of the causal estimate. By establishing causal relationships between weather and electricity consumption, policymaker and utilities stakeholders can make data-driven decisions regarding energy infrastructure and demand response programs with more conviction because a causal relationship is more strong than a correlation.

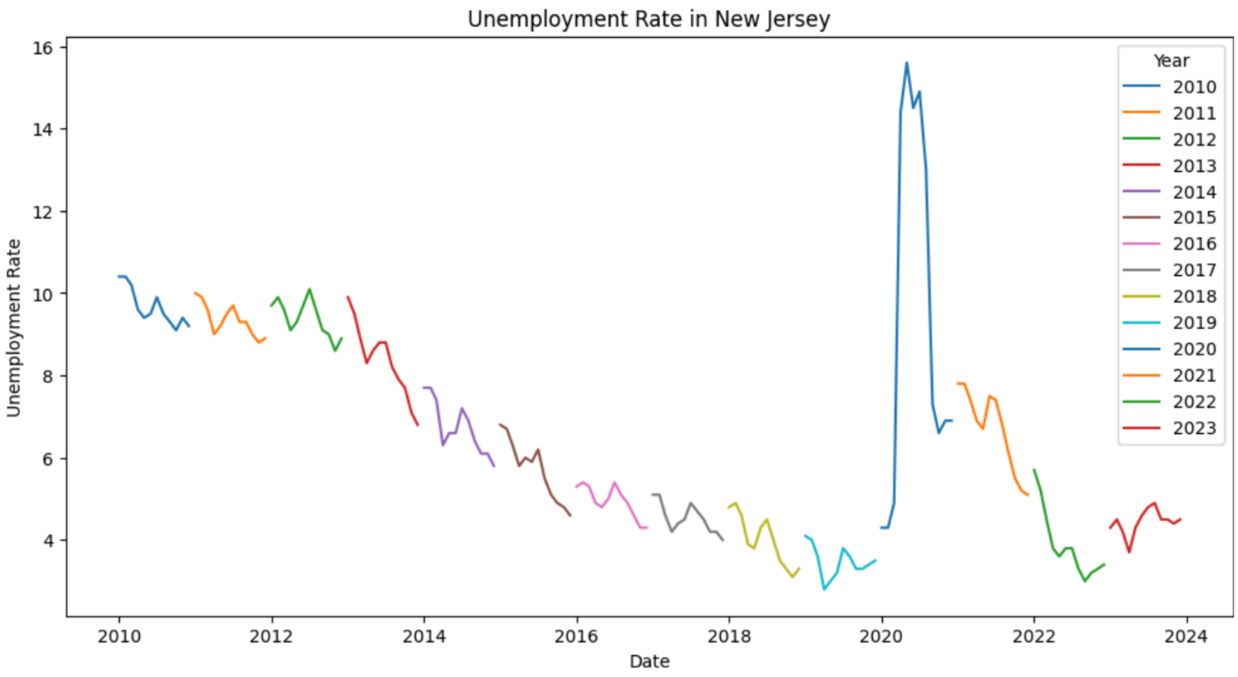
Even while accounting for confounding variables with propensity scoring, identifying and controlling for all confounders in this study can be challenging, leading to potential bias in estimating the true causal effect of weather on electricity consumption. Causal studies have a wide range of implications, and it is crucial that the study has integrity to ensure the public is not misinformed.

The second research question is **predicting electricity consumption in a region given the region's economic indicators**. The indicators we will consider are **unemployment rate, the consumer price index (CPI), and the Case-Shiller housing index**. Utility companies can use the predictive model to forecast electricity demand more accurately based on economic trends in a specific region. Utility providers already know how much supply they generate, so knowing demand is the missing piece to determine an optimal pricing strategy. In addition, accurate energy consumption prediction can help formulate policies that address climate change. A one-size-fits-all approach to fighting climate change is less effective than a nuanced plan that considers periods when the American economy is doing well or experiencing a downturn.

We will answer the second question using GLMs and a random forest regressor model. An advantage of GLMs is that they provide interpretable coefficients that help understand the direction and magnitude of the effects of economic indicators on energy consumption. We can assume that consumption increases when the economy is doing well because people have more disposable income to spend on energy consuming products. A GLM will allow us to dive deeper into this relationship and reveal what specific components of the economy (the economic indicators) have a positive or negative impact on energy consumption. We will also train and test a random forest model because it can more effectively capture complex relationships in the data when compared to other methods (especially in the case of non-linear relationships, which we suspect might be the case for our chosen variables).

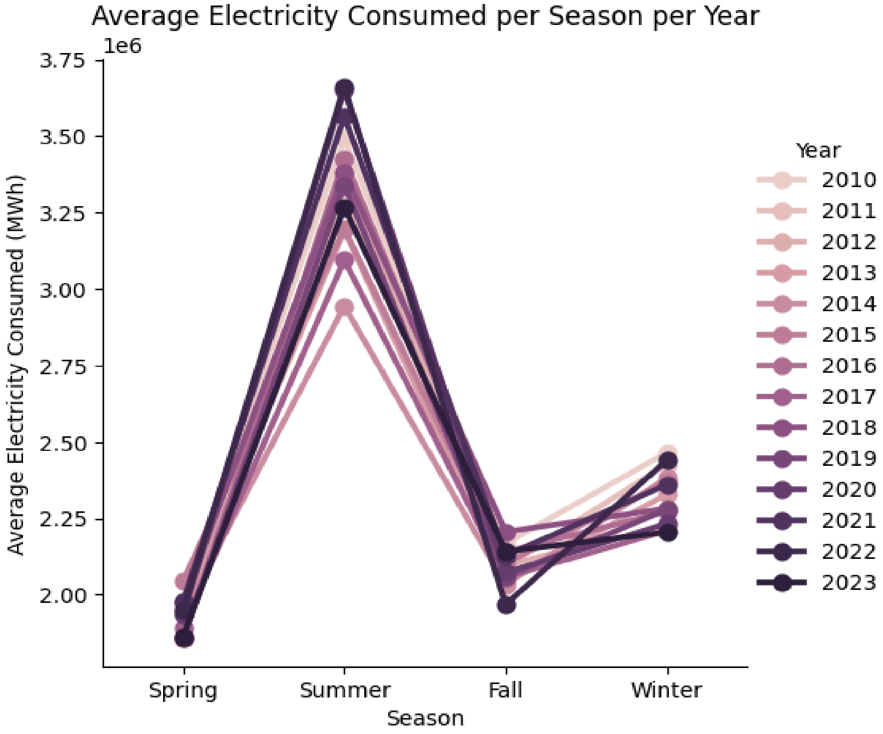
In contrast to random forest, the limitation with GLMs is that they assume a linear relationship between the economic indicator predictor variables and the consumption response variable. If the true relationship is non-linear, the GLM could have a low prediction accuracy. Additionally, a random forest model is significantly worse at interpretability than a single decision tree. Non-parametric methods are often considered black boxes that make it difficult to understand the decision-making process. Random forest models are also prone to overfitting, which is a fact we considered while conducting the study.

3. Exploratory Data Analysis



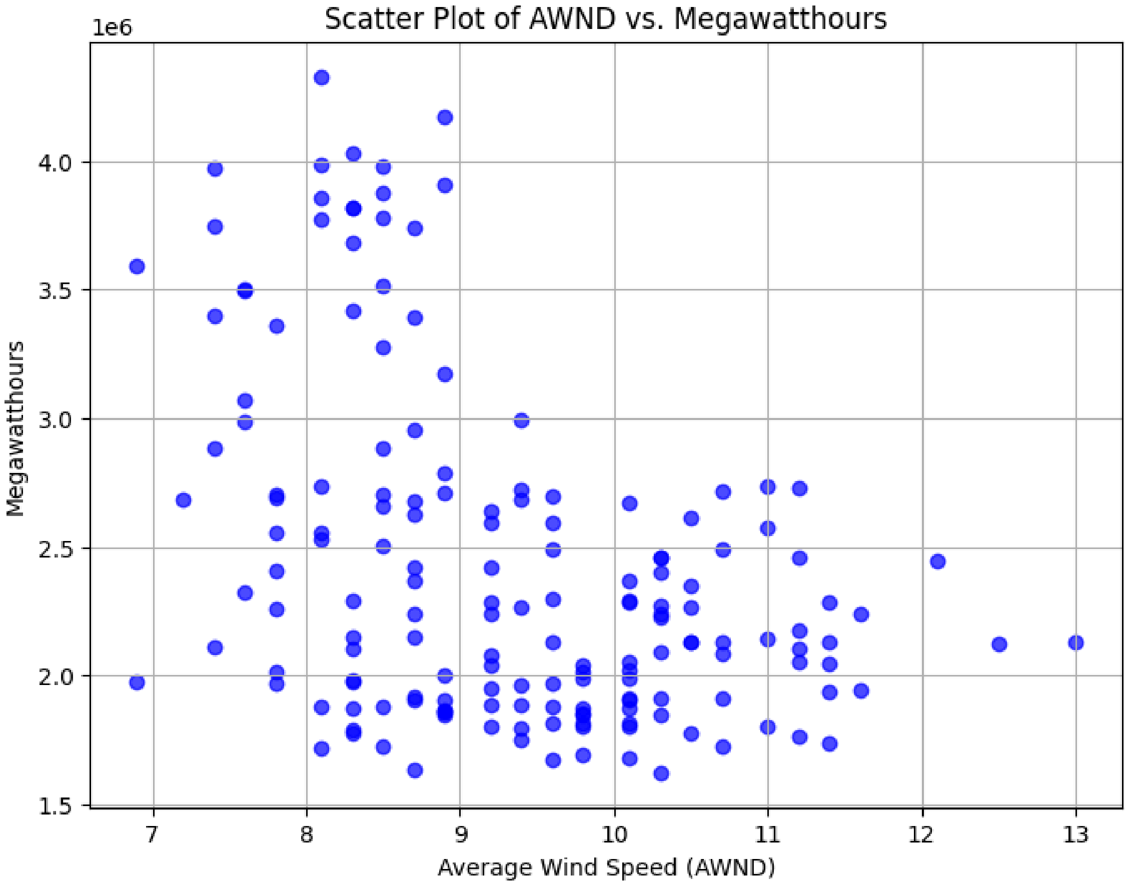
In our first graph, we visualize the changes in **unemployment rate** against **time (years)**. A few noteworthy points about this dataset: firstly, we see that there is a huge spike in unemployment during 2020. This is presumably because of the recessionary impacts of the pandemic, during which time an abnormally large number of people were laid off. Secondly, we also see that unemployment rates were steadily declining from 2010-2013 until 2020. This is again due to the economic conditions that were prevalent after the 2008 recession, so what we are seeing in the data corresponds to a recovery period. Post 2020-2021, we see that unemployment rates are starting to approach their pre-Covid levels. Another point to note is that unemployment rates tend to be falling across the year in which they were measured as well; for instance, nearly all years have higher unemployment rates in January of that year as opposed to December of that year (unemployment rate is measured monthly in our source dataset). Overall, the trends in the unemployment rate are expected, and show a pattern of decline barring economic aberrations like the pandemic and recessions.

The visualization here helps us answer our **second research question** as we are trying to predict how economic factors affect energy consumption. The graph essentially gives us insight into the health of the economy. This understanding of when and why economic headwinds and downturns occur can help us correlate to the relationships and differences in our data, in this case energy consumption. With these correlations, it can help us understand the true relationships between the impacts of the economy on energy consumption as a whole.



Our second graph visualizes the changes in **average electricity consumed** per **season per year**. In separating each year into seasons, we notice that firstly, the season with the highest average electricity consumption is summer, with the peak for all years being much higher than any other period. We also notice that there are declining periods of electricity consumption during transitions between summer and fall, and winter and spring. Another trend to note is that the yearly average electricity consumption (per season) seems to be increasing as we go from 2020 to 2023; we see that 2023 summer seems to be the highest average electricity consumption across the entire dataset. However, average electricity usage between 2010 and 2020, especially during the summer, seems to have been decreasing year-on-year, or at least lower than most of the years before it. Another interesting aspect about the data is that while summer electricity consumption is high (as expected), electricity consumption during the winter periods is not as high, although New Jersey winters are severe in their onset. This is interesting to observe as heating and cooling methods for residential housing may differ in origin sources for each in the state.

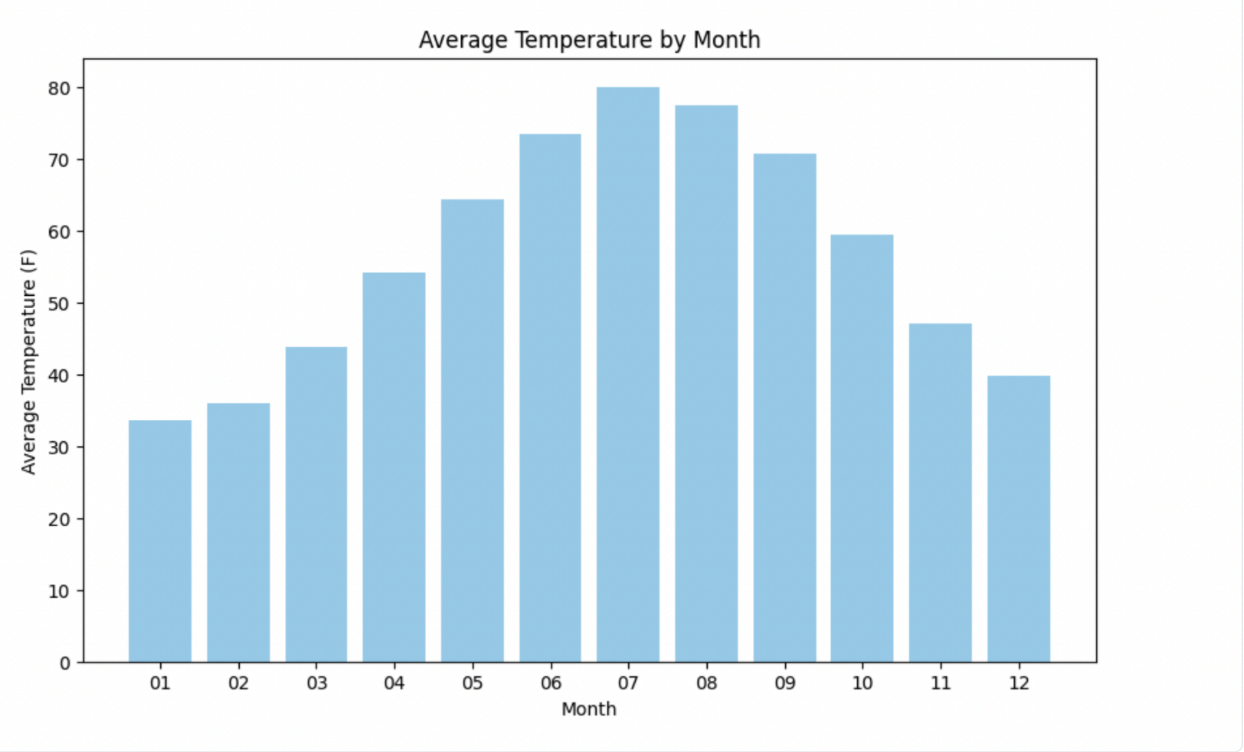
The visualization here helps us answer **both of our research questions** as it gives insight on the relationship between seasonal changes and the energy consumption. This relationship directly correlates to our first research question as the weather shifts over the seasons and with this weather shift comes changes in residential demand for energy. This visualization is also relevant to our second research question as seasonal and business cycles correlate heavily in terms of economic growth, unemployment and employment fluctuations, and high and low sales growth periods. Relating to seasonality, all of these economic factors in turn affect the energy consumption demand



This graph visualizes the relationship between **average wind speed** and **electricity consumption**. We can see that there appears to be some negative correlation between the two variables, as the data points are concentrated on or below the left diagonal. From the visualization, we can see that as the average wind speed increases, the electricity consumption seems to be decreasing, which is a result that is more or less expected. This seems to be a justified interpretation of this trend, because the upper right quadrant is nearly entirely devoid of any data points. An interesting trend in the data is that at a wind speed of 8-9, we see the highest usage of electricity (beyond 9, electricity consumption stays more or less within the same range). This is a relationship worth investigating further, because it could give us insight into potential confounders that may be causing this switch at this specific wind speed.

This scatter plot comparing Average Wind Speed to Megawatthours of energy consumed is relevant to our **first research question**, which investigates if there is a causal relationship between weather and energy consumption. In order to determine if there is a causal relationship, we first wanted to check if there is a correlation between the variables, which there seems to be.

There is a clear strong negative correlation between wind speed and megawatt hours. The reason for this relationship needs to be investigated further. This visualization gives us a good starting point to begin our causal analysis.



This graph visualizes **average temperature** from the yearly data measured over a **time on a monthly basis**. We notice that across the year (months 1 - 12), the average temperature increases then decreases, as expected, and that the highest average temperature occurs around July, which ties in with what we expect about a New Jersey summer. One thing to note is that apart from January and February, no two months are nearly identical in their average temperatures; that is, there are significant differences in average temperatures across nearly every month in New Jersey. Another interesting observation about trends in the data is that there is a considerably large range that is spanned by this dataset (around 50 degrees on average).

This visualization is relevant to our **1st research question** and encourages us to take into consideration the variation of temperature within a year and how this would affect our outcome variable, electricity consumed. Does higher average temperature lead to higher electricity consumption? What time of Does lower average temperature lead to higher electricity consumption? How different is the electricity consumption between months of high average temperature and months of low average temperature? These questions and visualization lead us to motivating the causal analysis on how weather affects electricity consumption (1st research question).

4. Research Question 1: *How does weather affect electricity consumption?*

Methods:

We are interested in three primary weather variables: temperature, average wind speed, and average precipitation. From this we have identified three treatment variables that we studied separately, described below:

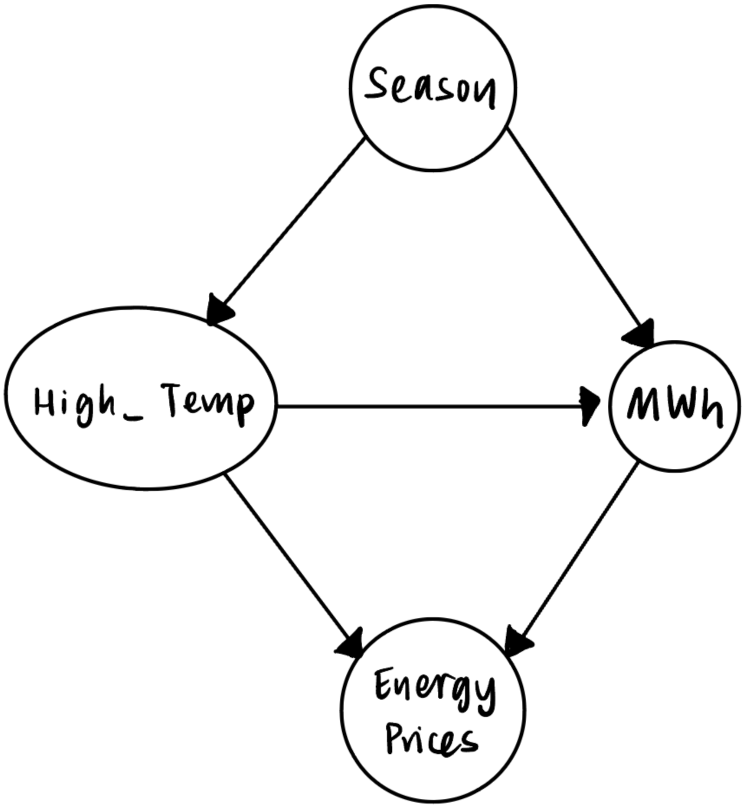
1. High_Temp: this variable is a dummy that indicates whether a given observed temperature is above the average temperature in New Jersey. Upon further research, we settled on using 53.6 degrees Fahrenheit as our threshold value in order to fully capture variation in the temperature in our dataset.
2. High_Wind: this variable is a dummy that indicates whether a given observed average wind speed (measured in miles per hour) is above the average wind speed in New Jersey. We calculated the average of these observations in our dataset as 9.33 and used this value as our threshold.
3. High_Rain: this variable is a dummy that indicates whether a given observed precipitation value is above the average precipitation in New Jersey. We again calculated the average of the observations in the dataset to use 3.98 (inches) as our threshold value.

We identified season as our confounding variable, because it is salient that the time of year will affect the temperature, and we found it reasonable to assume that the time of year will also affect electricity consumption through people's perceptions and preconceived notions about how hot or cold they expect to be in a given month. To adjust for seasonality, we are using our propensity score analysis and we have accounted for season in our logistic regression.

We identified the price of electricity as a potential collider, because the consumption of electricity affects prices through supply economics, and the temperature also influences electricity prices through demand and surge pricing.

The causal DAG for these relationships can be captured as follows:

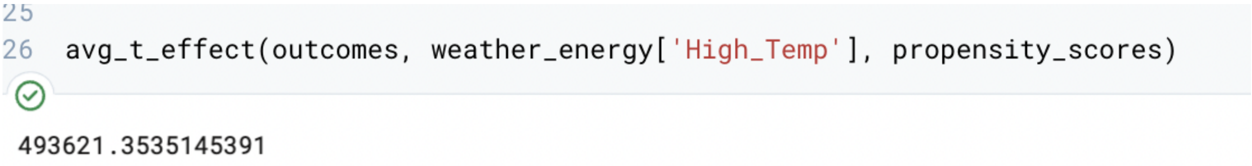
(MWh = megawatt hours; units of electricity consumption)



Results:

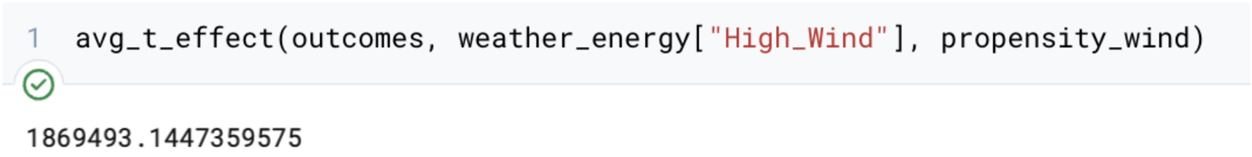
We performed a propensity score analysis. We first created the binary variables High_Temp, High_Wind, and High_Rain, and then ran logistic regression models for each of these variables. From the coefficient estimates generated through the regression, we obtained propensity scores and then applied the inverse propensity weighting formula below (lecture slides, Strang and Sridharan):

$$\hat{\tau}_{IPW} = \frac{1}{n} \sum_{i:Z_i=1} \frac{Y_i}{e(X_i)} - \frac{1}{n} \sum_{i:Z_i=0} \frac{Y_i}{1-e(X_i)}$$



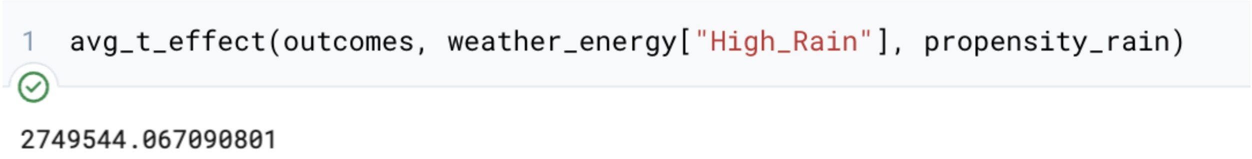
```
avg_t_effect(outcomes, weather_energy['High_Temp'], propensity_scores)
```

```
493621.3535145391
```



```
avg_t_effect(outcomes, weather_energy["High_Wind"], propensity_wind)
```

```
1869493.1447359575
```



```
avg_t_effect(outcomes, weather_energy["High_Rain"], propensity_rain)
```

```
2749544.067090801
```

When looking at each of the ATE's present here between each of the factors and the outcomes, we note that all of them are positive. The average treatment effect can be interpreted as the difference in mean electricity consumption between the treatment group (observations where either temperature was above 53.6 degrees Fahrenheit, average wind speed was above 9.9 mph, or average precipitation was above 3.98 inches) and the control group (observations in the control), and because this difference is so large, this means that High_Temp, High_Wind, or High_Rain are all expected to have a positive relationship with electricity consumption. With High_Temp having the highest average treatment effect, we can conclude that the high temperatures factor has the most substantial impact when it comes to residential energy demand at the rate of the increased energy consumption. We can also conclude that the relationship we have found is causal.

We have assumed conditional independence between the treatment (High Temp, High Wind, and High Rain) and the potential outcomes. In other words, we have assumed that given the propensity scores, there are no unmeasured confounders affecting both treatment and outcomes. We know this because we have selected covariates that have been previously found to be associated with treatment and outcome, which alludes to our conditional independence.

However, the Average Treatment Effect estimates for each treatment are not statistically significant. We used a bootstrap (p-value calculation) and found that p-values are all above 0.05. In other words, we fail to reject the null hypothesis that there is no causal effect between the treatment to the electricity consumption (ATE = 0). Therefore, we cannot conclude that the treatments (high temperature, high wind, and high rain) have a statistically significant causal effect on electricity consumption.

Discussion:

One limitation on our methodology is our selection of the threshold values for defining High_Temp, High_Wind, High_Rain. As they were calculated using the averages within our dataset and respective timeframe, the threshold may not capture the full variation of weather conditions in the state of New Jersey. Additionally, along with seasons which we accounted for, there can be other confounding variables that may need to be considered as they can affect the relationship we view in our research question. Additional data that would be useful would be data including the other potential confounding variables and also interesting data like the state energy infrastructure so we can analyze the infrastructure limits of New Jersey and how it can handle fluctuations in the energy demands through the year. As mentioned previously, as our ATE estimates are not statistically significant, we cannot definitely conclude that there is a significant causal relationship between our three treatment variables and electricity consumption.

5. Research Question 2: *How do economic indicators (CPI, unemployment percentage, Case-Shiller Housing Index) influence electricity consumption?*

Methods:

In this research question, we are trying to predict electricity consumption (Megawatt Hours) from economic indicators such as the CPI (Consumer Price Index), the Case-Shiller Home Price Index, and the unemployment percentage using general linear models and nonparametric models.

1. The CPI, a measure of the average change in prices for a standard set of goods and services, plays a crucial role in understanding electricity consumption. As one of the essential utilities a household uses, electricity is influenced by changes in the CPI. These changes can result in more purchasing power, increasing electricity consumption, and vice versa.
2. The Case-Shiller Home Price Index, a set of indices that measure the monthly change in residential real estate prices, is a relevant economic indicator for our study. As we focus on electricity consumption from residential households, a change in the housing index can indicate increased wealth and, consequently, more purchases of electricity-consuming products.
3. Unemployment percentage is the number of unemployed people divided by the number of employed people times 100. Unemployment percentage has a more behavioral effect on how people use electricity. Higher unemployment rates tend to lead to a more conservative use of electricity for several reasons. An unemployed person may want to conserve energy to avoid paying high electricity bills. When job security is unstable, people may have a pessimistic view of future economic prospects that they've adopted a conservative lifestyle. Those are a few ways the unemployment percentage can affect electricity consumption.

The GLM we have decided to use is the Log-Normal GLM. This is because Electricity Consumption (Megawatt Hours) is continuous with non-negative values, and the distribution is right-skewed. When the log function is applied to the outcome variable (Megawatt Hours), it turns into a normal-looking distribution. The Log-Normal distribution has two parameters, mean and variance, making it easily applicable to overdispersed data such as electricity consumption.

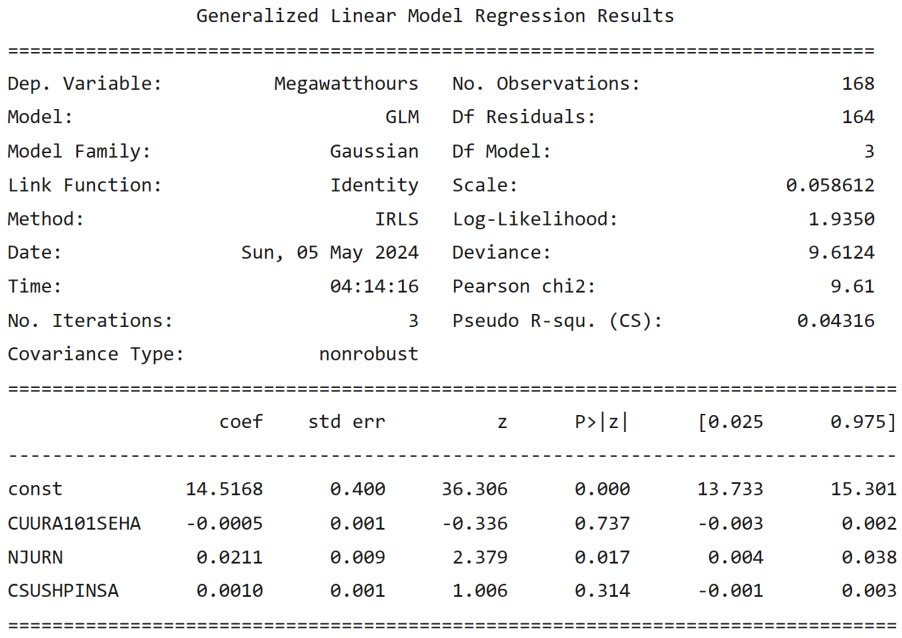
With the Log-Normal distribution, we make assumptions that each instance of the dataset is independent of each other and that there are no outliers since the log-normal distribution is sensitive to outliers but less so than the normal distribution. We also assume that each predictor variable (CPI, Case-Shiller Index, and unemployment rate) has a linear relationship with the outcome variable (electricity consumption), which we later find false. The relationship is linear in the condition of season and year. However, since we are only looking at a relationship between economic indicators and electricity consumption, we could not include those variables, however important they are. We also assumed that we needed some vital predictor variables, such as GDP, but could not retrieve that information due to a lack of records prior to 2018.

As for the non-parametric version, we have decided that the random forest would be our best model to predict electricity consumption based on CPI, Case-Shiller Housing Index, and unemployment rate. Random forest is best fit to model the relationship due to its versatility and ability to capture relationships between the predictors. Since all three predictors are economic variables, there is bound to be some type of correlation that they have. The assumption we made with the random forest is that the aggregated prediction of the decision trees is better than the individual trees in terms of variance. We also assume that the random forest can capture the non-linear relationship between predictors without the need for feature engineering or transformation.

To evaluate each model's performance, we will assess the log-likelihood and chi-squared values for the GLM model. For the nonparametric model, we will use the mean squared error to evaluate the model's performance.

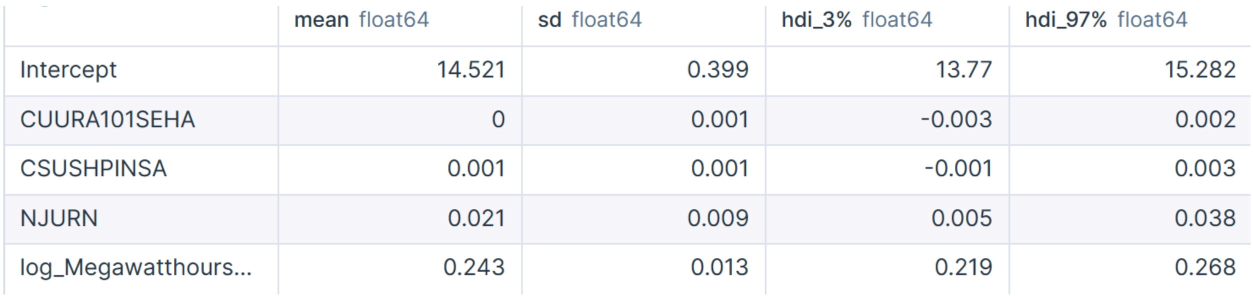
<u>Results:</u>

Log-Normal GLM (Frequentist Approach) -

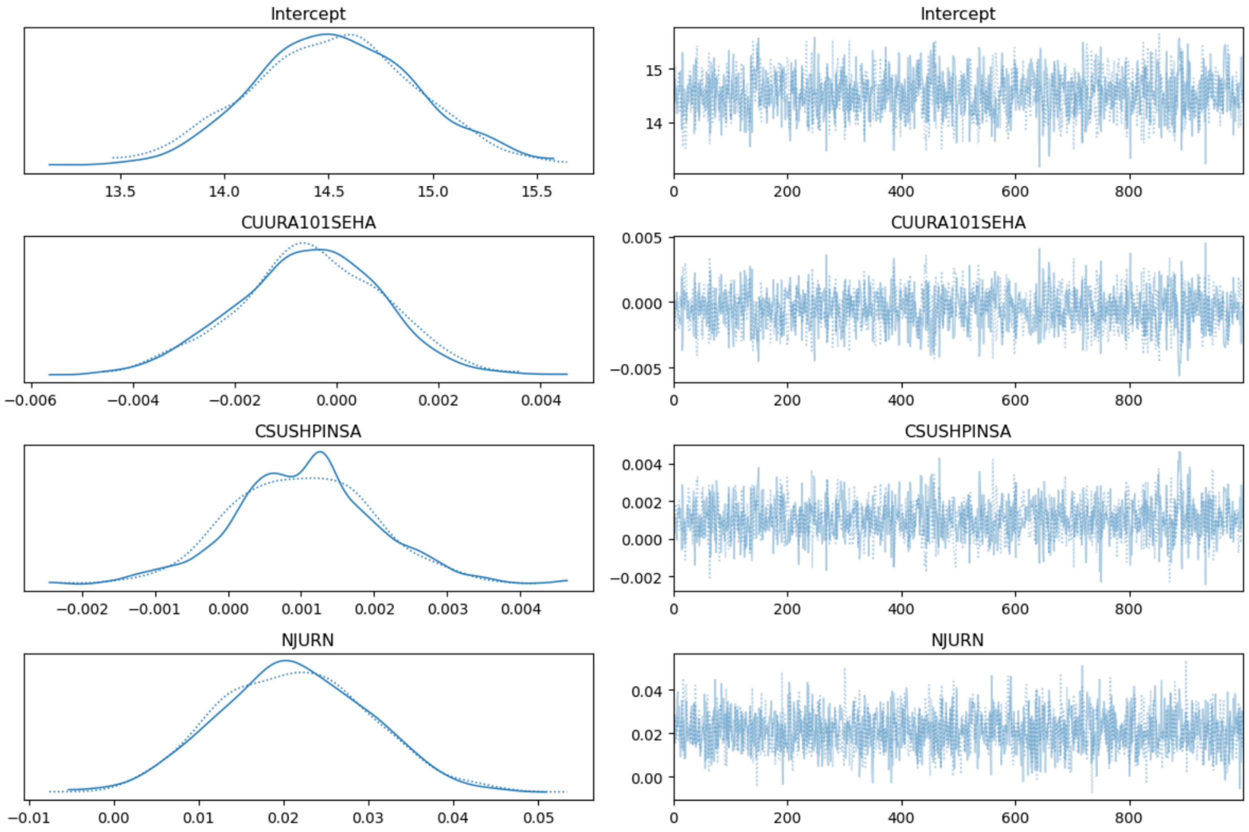


```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:          Megawatthours   No. Observations:                  168
Model:                            GLM   Df Residuals:                      164
Model Family:                Gaussian   Df Model:                            3
Link Function:               Identity   Scale:                        0.058612
Method:                          IRLS   Log-Likelihood:                 1.9350
Date:                Sun, 05 May 2024   Deviance:                       9.6124
Time:                        04:14:16   Pearson chi2:                     9.61
No. Iterations:                     3   Pseudo R-squ. (CS):            0.04316
Covariance Type:            nonrobust
================================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
const         14.5168      0.400     36.306      0.000      13.733      15.301
CUURA101SEHA  -0.0005      0.001     -0.336      0.737      -0.003       0.002
NJURN          0.0211      0.009      2.379      0.017       0.004       0.038
CSUSHPINSA     0.0010      0.001      1.006      0.314      -0.001       0.003
================================================================================
```

Log-Normal GLM (Bayesian Approach) -



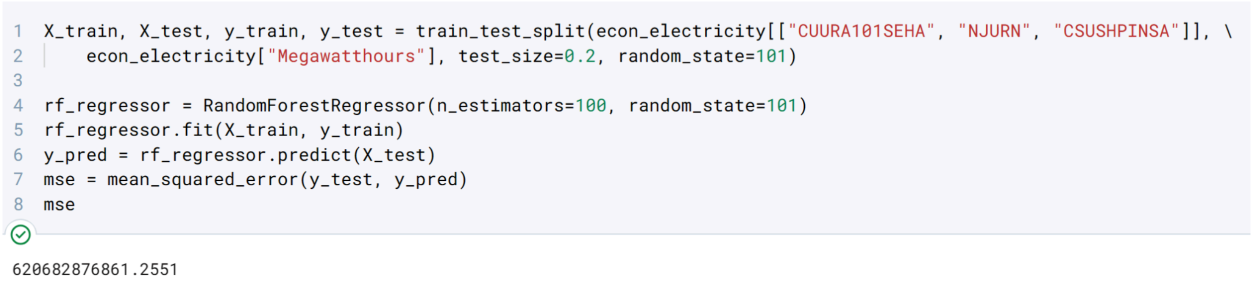
	mean float64	sd float64	hdi_3% float64	hdi_97% float64
Intercept	14.521	0.399	13.77	15.282
CUURA101SEHA	0	0.001	-0.003	0.002
CSUSHPINSA	0.001	0.001	-0.001	0.003
NJURN	0.021	0.009	0.005	0.038
log_Megawatthours...	0.243	0.013	0.219	0.268



Sidenote:
CUURA101SEHA = Case-Shiller Home Price Index
NJURN = Unemployment percentage
CSUSHPINSA = CPI

Based on the coefficients above, the Case-Shiller Home Price Index, Unemployment percentage, and CPI have a very small effect on electricity consumption, with the unemployment percentage having the most effect. A one unit increase in unemployment percentage results in an exp(0.0211) decrease in electricity consumption (Megawatt Hours). This makes sense based on the justifications made about unemployment's effect on electricity consumption. Out of all the predictor variables, the unemployment percentage does not have a confidence interval or credible interval, including zero in its interval. So, the unemployment percentage does have a real albeit tiny impact on electricity consumption, while CPI and the Case-Shiller Home Price Index have no real effect on electricity consumption.

Nonparametric Model - Random Forest:



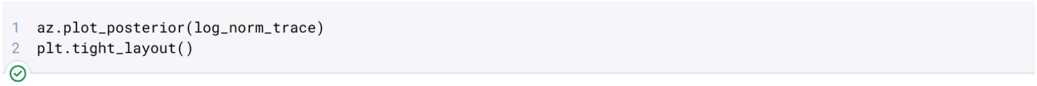
```
X_train, X_test, y_train, y_test = train_test_split(econ_electricity[["CUURA101SEHA", "NJURN", "CSUSHPINSA"]], \
    econ_electricity["Megawatthours"], test_size=0.2, random_state=101)

rf_regressor = RandomForestRegressor(n_estimators=100, random_state=101)
rf_regressor.fit(X_train, y_train)
y_pred = rf_regressor.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
mse
```

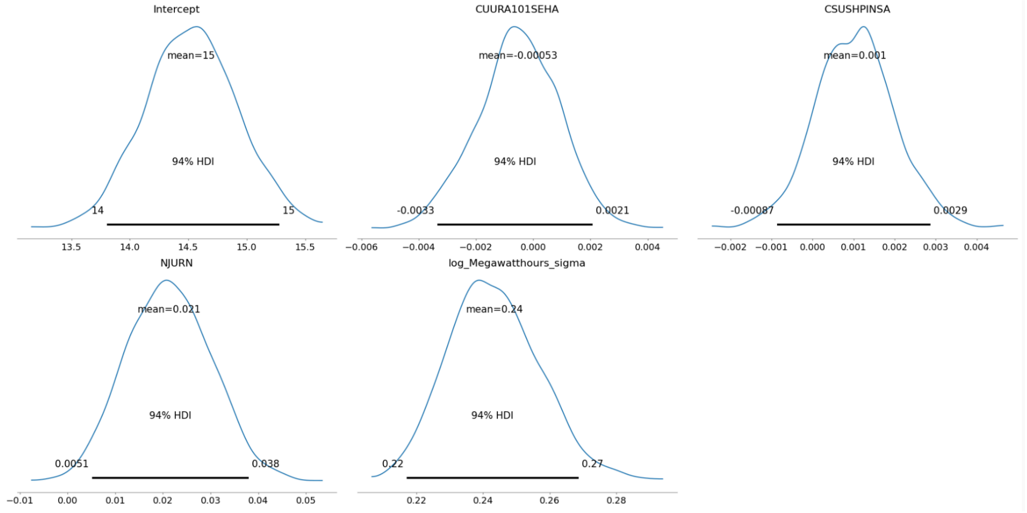
```
620682876861.2551
```

Based on the large mean squared error value, the model does a poor job of predicting electricity consumption. This may be because the predictor variables do not have a direct effect on electricity consumption. In other words, our model is underfitting due to the lack of explanatory variables.

Uncertainty in the GLM Prediction



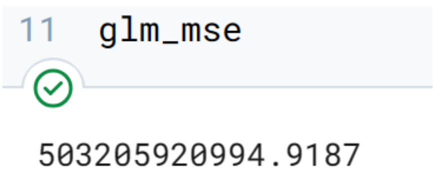
```
az.plot_posterior(log_norm_trace)
plt.tight_layout()
```



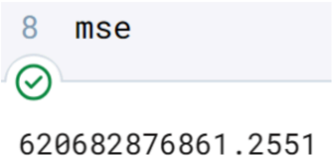
The posterior distribution of log_Megawatthours has an HDI of [0.22, 0.27] and a mean of 0.24. This means we are 94% sure that the Credible Interval, [0.22, 0.27], contains the true parameter of the log_Megawatthours given the data. The same goes for each parameter and their respective interval.

<u>Discussion:</u>

MSE for Predicting Unseen Data



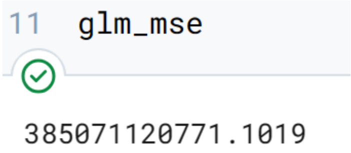
```
glm_mse
503205920994.9187
```



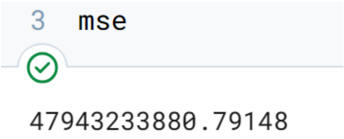
```
mse
620682876861.2551
```

Left - Log-Normal GLM
Right - Random Forest Model

MSE for Fitting the Training Data



```
glm_mse
385071120771.1019
```



```
mse
47943233880.79148
```

To assess which model performs better, we trained and tested both models and evaluated their mean square error for both the training and testing datasets. Overall, the GLM model does a better job of predicting unseen data and fitting the training data. However, both models perform poorly in predicting and fitting the data, as indicated by the large MSEs. One possible reason for the GLM model's better performance compared to the non-parametric model could be that the predictor variables have a more linear relationship with the outcome variable. Additionally, unsupervised models are typically used to address overfitting, but in this case, we do not have an abundance of predictor variables, resulting in underfitting of our model. Based on these MSE results, we cannot confidently use either model on future datasets, as there is too much discrepancy between the predicted values and the true values, making them no better than a random number generator.

6. Conclusion

From our causal question analysis, we found high ATE values for all three of our variables of interest (above average wind speed, temperature, and rain) indicating that the average electricity consumption in an area with above average weather conditions will on average be significantly higher than the average electricity consumption in an area without, but these results are not statistically significant. From our analysis in our second research question regarding economic indicators and their relationship with electricity consumption, we found that our random forest method had a large mean squared error value, indicating that the economic indicators may not be directly related to our response variable, but our GLM analysis shows that while CPI and the Schiller index may not have a statistically significant impact on our response variable, the unemployment rate is negatively related to electricity consumption and this relationship is statistically significant.

Because of our choice of state (New Jersey), our findings about weather may be limited to states that are situated in the northeast, with similar weather patterns, but as noted in our data overview, the average weather conditions in New Jersey are on average similar to the rest of the country, save for extremes. Our economic indicator data is less generalizable because it was very specific to the state; however, economic indicators tend to move together, and during our time period of study, the entire country went through two recessions and so unemployment data, CPI data, and housing prices would have been roughly moving in the same direction. One potential next step would be to carry out a difference-in-differences approach to test this claim, and see exactly how generalizable our findings regarding the relationship between economic indicators and electricity consumption are.

In light of our findings, we can state that if New Jersey officials want their electricity consumption to increase in order to boost output and productivity, perhaps, then they should target expansionary fiscal and monetary policies that would lower the unemployment rate. Similar studies should be run to see if manufacturing electricity consumption depends on economic indicators (instead of just the residential analysis which we performed), which would then suggest more directly that decreasing the unemployment rate would increase manufacturing and perhaps even bolster the state's balance of payments.

Through this project, we uncovered relationships in our data by applying several quantitative and analytical methods. We learned about the nuances of data exploration and what a research project cycle actually looks like. Combining different sources of data allowed us to explore more holistically the potential relationships between our variables of interest, as well as ensuring that we didn't run into omitted variable bias. However, we only looked at New Jersey data, and this could have led to some of our results not being statistically significant. Future studies could take on a larger collection of data and apply the same methods, and include more explanatory variables in the analysis.