Sarah Patten
Energy Consumption: Milestone 1 Report

Problem Statement:
The United States consumes various forms of energy every day. Energy companies would be interested in seeing how different economic data affects energy consumption. If they know how the economy is trending with GDP data, unemployment rates, or what state / region they're dealing with, they can better adjust their prices to obtain the biggest return.

The Data:
For this project, a few different datasets were used. The first is a CSV file with 52 rows and 192 columns. Each row represents a state, and each column breaks down different economic aspects as well as energy consumption. This data came with a few challenges; the different years weren't their own rows, as only states were determining factors for rows. The columns also had energy broken down by type, consumption, production, expenditure, and price for each year.

To be able to use these columns, I first searched for column names contain string versions of the years. I then had lists of columns names for each year, and another group that didn't contain a year (these are the base columns that are needed like state, region, whether the state borders a coast, state codes, and whether the state border one of the great lakes. Each list of years columns was combined with the list of "other" columns, and that list of column names were selected from the original dataset, and a new "Year" column was tacked on the end. When the year columns were broken out into their own lists, the "year" was dropped from the end of the column names. I then concatenated all of the tables for each year back together, since they all had the same column names, they were just stacked on top of each other.

Once the data was broken out by year the CENSUSPOP column was dropped, due to the large number of missing values. After that the process that was done for each year needed to be redone, for each energy type this time. Energy columns had the energy type as the start of the energy column, with either C, E, P, or Price at the end of it. So I repeated all the steps from the "Year" process, for the energy types, and created the new column for "Energy_type", and renamed the "C", "E", and "P" columns with "Consumption", "Expenditure", and "Production".

The other datasets that were used were Unemployment and GDP per capita tables. The two CSV files were joined to the first dataset, prior to breaking it out by year and energy type. The first thing I did was drop the year values from the table that weren't available in the main table, then a few columns that wouldn't be used. I then consolidated the unemployment data by grouping computing the average unemployment rate for a state, for each year (previously there was a row for each county within a state). Next the data had to be pivoted so that each row represented a state, and there was a different column for each year. Once this was done we

had close to the same shape as the main data, so they were left joined by state. There ended up being a few states that were missing from the unemployment data that would be handled later.

The GDP per capita went through roughly the same transformations as the Unemployment table before it was able to be joined to the main dataset. However, this table actually had some extra “states” that were not available in the main table. But, because they were actually regional, and the data was again left joined, with the main dataset on the left, these excess columns were lost.

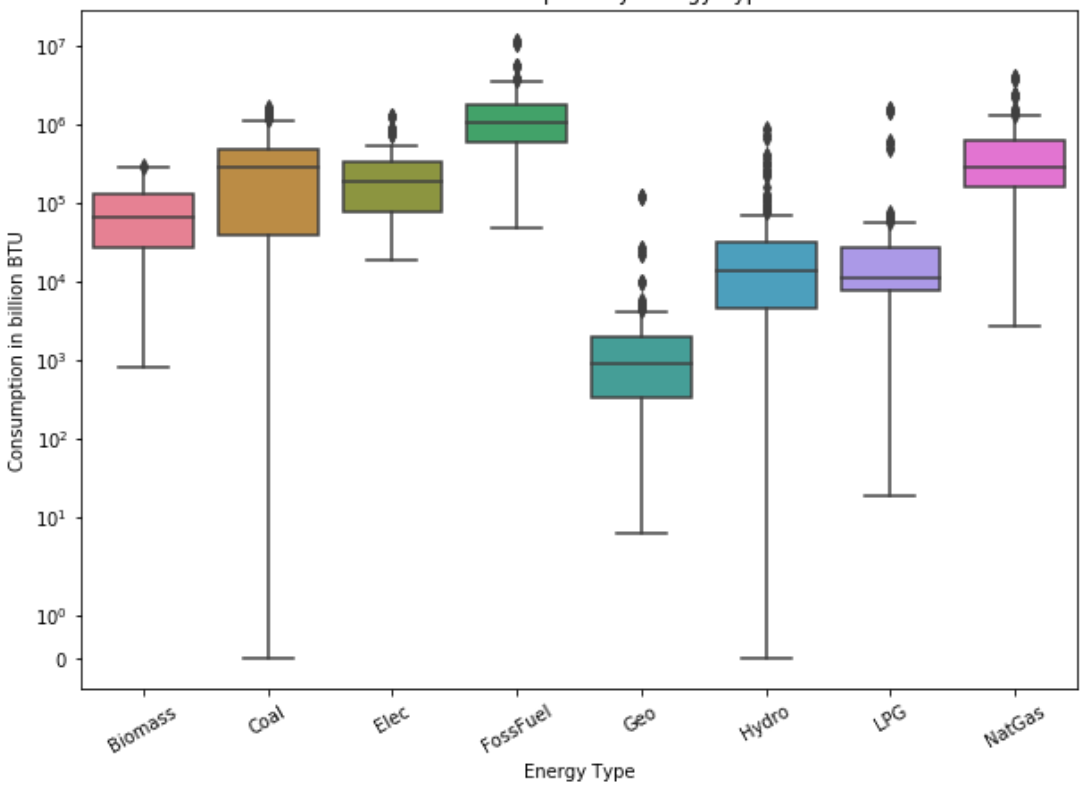
Now that all the data was in 1 table, the missing information needed to be cleaned up. For most of the columns, the average of the column sufficed for filling in the missing data, but for a couple, the missing data represented things that should actually be considered 0, so that’s how those columns were filled. But having that many zeros throws off the Price data a little.

Seven of the columns; State Codes, States, Energy type, Region, Coast, Division and Year were all converted to categories as there were a limited range of values for each of them.

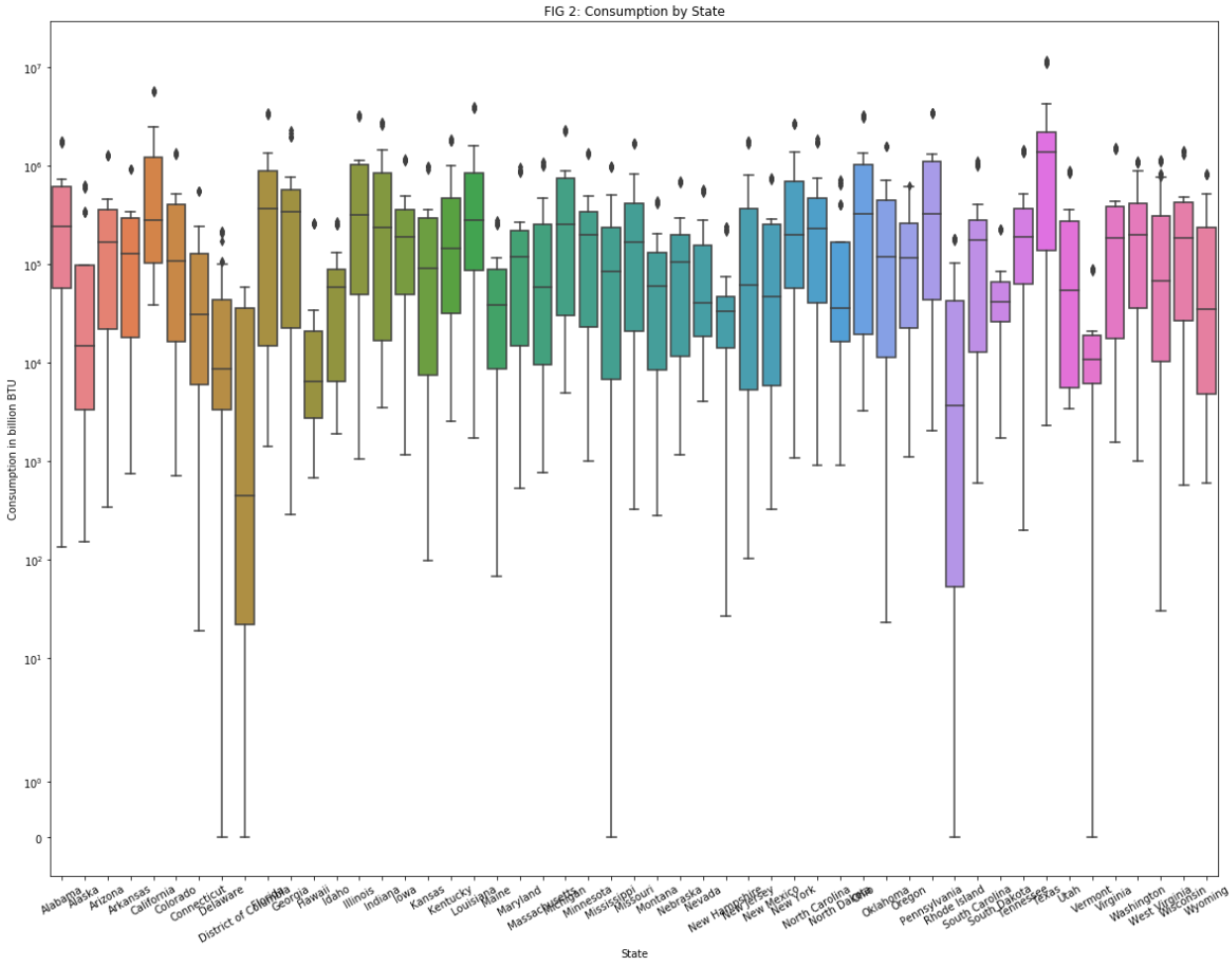
Data Visualizations:

The first thing that felt important to look at was the spread of consumption information by each energy type. The boxplot is shown in Figure 1. Wit it we can clearly see that Fossil Fuels are the most used energy type, and Geothermal is the least used.

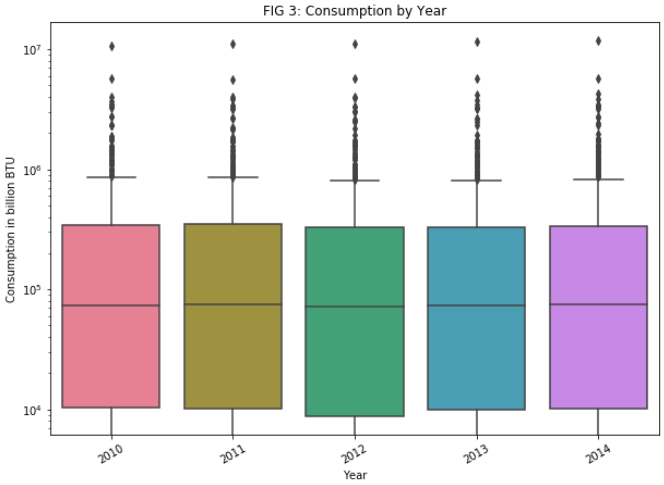
FIG 1: Consumption by Energy Type

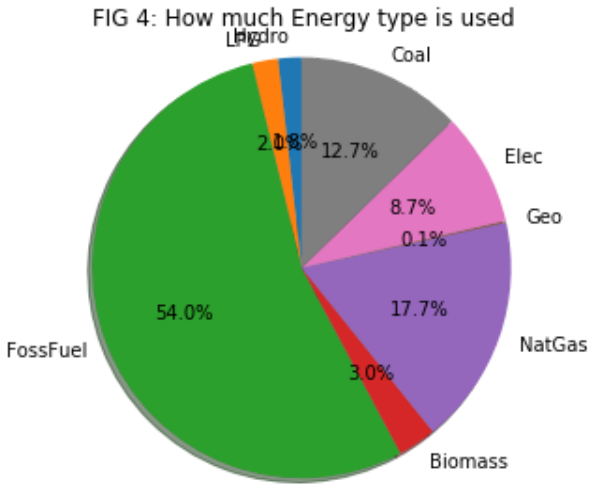


Plotting out the energy consumption by each state, regardless of the energy source was interesting to see. While it's a lot of data, we can easily see the high point was for Texas. They have the largest outlier, as well as highest average energy consumption. On the other extreme, Delaware has the lowest average and one of the biggest spreads of data.

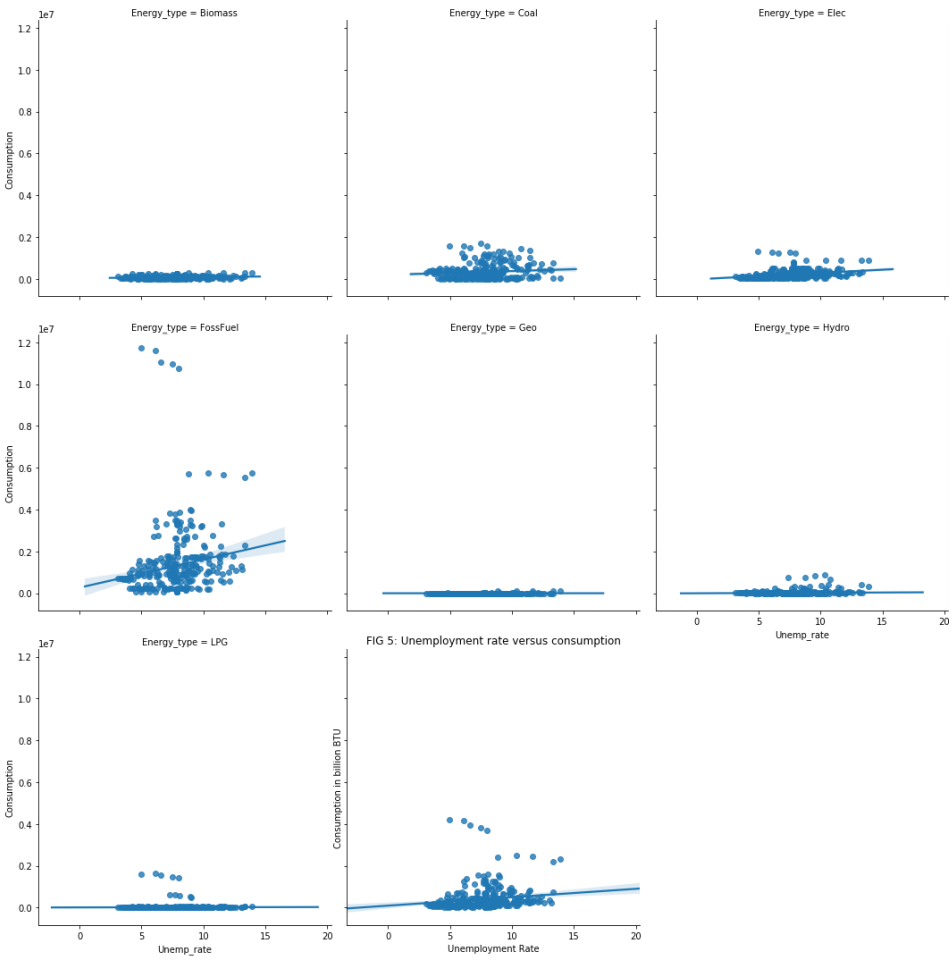


The yearly data shows there isn't much of a change in the consumption amounts between the years; there seems to only be a very slight upward trend for energy consumption by year, with minimally larger outliers. But, because it's so minimal, it could just be a trick of the eye from the large scale of the data. It is a little disappointing that it doesn't show as wide a range of variability as the other categorical variables.



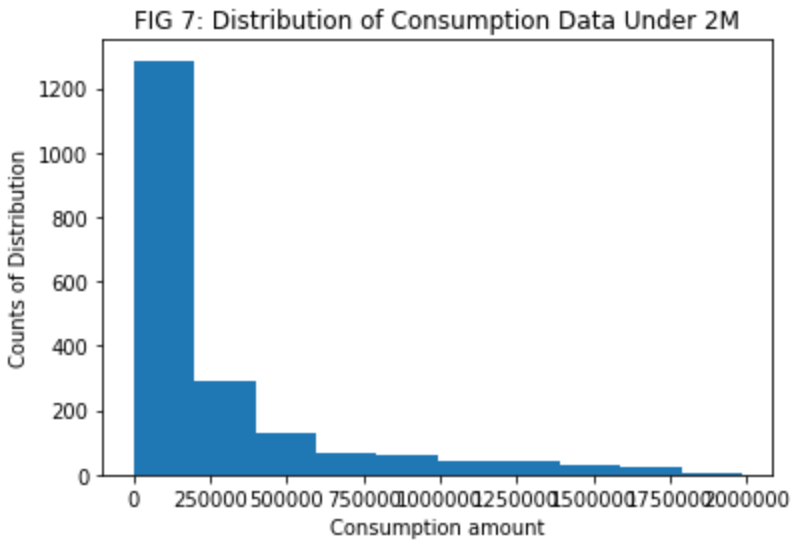


As shown in the first boxplot we could tell that fossil fuels were the most largely used of the energy types, the Pie Chart in Figure 4 shows that it actually accounts for 54% of the total energy usage for all the years in the dataset, Geothermal accounting for only 0.1% of it, and Electric for only 8.7%. The large amount of fossil fuel usage makes sense with the number of cars on the road and the amount of driving done every day in the United States. While it isn't shown in the years of data the transition to electric cars may help to change the percentage to be more electric usage, and less fossil fuel each year.



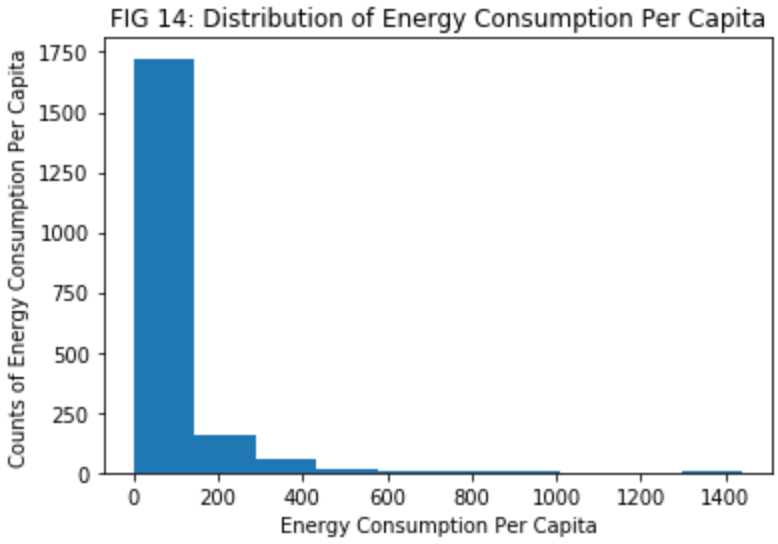
These 7 charts show the influence of the unemployment rate against the Energy consumption. Most of the graphs don't show much of a correlation, probably due to the small amount of consumption for those types. However, the Fossil Fuel, Electric, and Natural Gas show that the higher the unemployment rate, the more energy is consumed. This is because the more people are at home instead of at work, the more they're using electricity, whether that be through home computers, televisions, or air conditioning, while the businesses would use close

to the same amount of energy no matter the employees, because they still keep the temperature roughly the same, and the electrical equipment in an office or building would still be running approximately the same.



The total distribution of consumption data is displayed in Figure 7. Of course, the high end outliers were dropped prior to the graph being plotted. In this, we can see that the data is exponentially distributed. The consumption amount is in billions of BTU. Since we had broken the energy down to be by energy type, and year, it makes sense that some years, close to 0 billions of BTU were used for say, geothermal energy. The consumption data is the most interesting portion, because it's what energy companies will be looking at. The next thing to be looked at is the energy consumption per capita.

To format this data, we needed to create a new column that encompassed the Consumption data divided by the population estimate. Because the Consumption data was in Billions of BTU, to get it into millions, we simply needed to multiple by 1000. Now we're in a readable scale. This information is still exponentially distributed, so we won't be able to use some statistical analysis methods to approach it.



Statistical Analysis:

Reading through the U.S. Energy Information Administration (E.I.A.) I found that they reported the average consumption per capita, of the primary energy source, in 2016 as 302 million BTU. The average BTU usage per capita from this dataset was calculated as 314 million BTU per capita. There is a slight discrepancy between the two, so by pulling some bootstrap samples, we'll be able to test how accurate our dataset is.

Because the E.I.A. only looks at the primary energy source, we grouped the data down to find out which energy source had a max value for each state and year, then kept that row's data. This is what's plotted out in Figure 14.

The majority of our data falls below 500 million BTU per capita, with an average of 314.33 million BTU. If we drop values above 500 million BTU per capita, the average drops to 246.83 million BTU per capita. We will hypothesize that the true average Consumption per Capita is 302 million BTU.

- H0: mean(Consumption per Capita) = 302 million BTU
- H1: mean(Consumption per Capita) ≠ 302 million BTU

Using the full data, bootstrap samples are drawn to see if our mean found from our data is true and the mean listed on E.I.A is wrong, or if the deviation from the mean is within reason. We'll use bootstrap samples to determine this with a p-value of 0.05, following the assumption that the mean is actually 302 million BTU and our dataset is a little off.

The first step of doing this is scaling our data to have a mean of 302, instead of 314. We do this by subtracting the mean from each row of Consumption per Capita, and then adding 302 back to the column. After this is done, we draw 10,000 samples and keep a count of how many of these sample resulted in a mean of 314 or larger. It turns out that we end up with a sample mean of at least 314 nearly 20% of the time (our p-value for this test was 0.2085), making it reasonable to accept our null hypothesis that the true mean consumption per capita is 302 million BTU.

Predictions of consumption:

The most interesting part of this data is to attempt to predict how much energy will be consumed. There are five years of data being analyzed in this project, so for out predictions we'll use the first four years to make predictions for the fifth year. We'll start by doing some standard regression modeling; linear regression, lasso, and ridge, followed by a couple more complex models with Random Forest and K Neighbors. Because we have so many columns (predictors) in our data, we should do pretty well with our predictions.

The first three, simpler, models will be used just to get a feel for how well we can predict the data. We'll start with default hyperparameters, then later tune them to see if we can obtain better results.

	r2_train	r2_test	mse_train	mse_test
LinearRegression	0.544138	0.53623	3.20852e+11	3.48251e+11
Lasso	0.544112	0.536235	3.20871e+11	3.48247e+11
Ridge	0.544086	0.536131	3.20889e+11	3.48325e+11
RandomForestRegressor	0.991565	0.985716	5.93674e+09	1.07262e+10
KNeighborsRegressor	1	0.548655	0	3.38921e+11

The goal is to predict energy consumption per capita.

Our Standard regression models didn't perform well at all. We couldn't even predict

at 55% with any of them. Why is this? There are over 30 predictors in this analysis, and this is probably causing some extreme overfitting. We can see in even our default random forest that we need to use fewer predictors at a time. The default random forest gave us a score of 97.8%! That's something that doesn't even require further tuning. But, just to see if we can do any better, the parameters of the random forest regressor were tuned to build a better model, and we were able to get a score of about 98.7%.

Of the models, this was the best, and we were able to see how important Price was in predicting energy consumption! Surprisingly, unemployment didn't fall in the top 10 for predictors, but the type of energy being predicted was incredibly important; having two of the eight energy types in the top 10 for predictors. Most of the states weren't good predictors, except Texas, which was shown to have significant energy consumption earlier.

Conclusions:

What this all means; energy companies will now be able to tell with reasonable certainty how much energy a person will consume, depending on what type they're supplying. Because they know price is a large factor in predicting how much will be used, the energy companies can control what type of energy people will consume, simply by changing the pricing of them. The more expensive they make a form of fuel, the less people will use it. This is something California has seen success with in the past. They're currently using this idea to force people towards using full electric vehicles. They have been upping gas prices and tax on gas to force people to stop driving their cars as much, and either switch to an electric vehicle, or start using more public transportation. They have also begun setting up more free charging stations. This gives consumers negative incentives to keep their current vehicles and a major positive influence on buying electric vehicles.

The consumers can use this information as well to understand why energy companies may be altering their prices. By figuring out what energy sources are trending, or dropping in prices, they can get the most out of their money but switching to a new fuel.