United States GDP Exploration of Data Analysis 2010 - 2014

Submitted by:

Donald Kane, Graduate Student at Vanderbilt University
Anubha Nagar, Graduate Student at Vanderbilt University
Xinyu Gao, Graduates Student as Vanderbilt University
Zhengqi(Peter) Tian, Graduate Student at Vanderbilt University

7 December 2021
Dr. Michael Shepherd

DSI 5610

United States GDP Exploration of Data Analysis, 2010 - 2014

Introduction

The COVID-19 pandemic had a devastating effect on the GDP of the United States. Due to the increase in bankruptcies, unemployment and layoffs the capital in the hands of the people dwindled. This caused the GDP to fall at rates we could never have imagined. In this paper, we will analyze past data from the years 2010-2014 to understand what are the driving features for the GDP of the United States. We have analyzed the energy, census and consumption data to determine which parameters affect the GDP. We have also done an in depth analysis about which sort of energy(i.e. Geothermal, coal, electricity etc) affects the gdp, which census data/ consumption data affects the GDP. Moreover, we will justify our statements with graphical explanations to better justify our statements.

Data collection

The data was collected by the US government on census, geographic, energy and economic data and is available on Kaggle. The dimensions of the data are 52 rows and 192 columns. The distribution of columns is as follows:

- 30 Columns for Census data (includes Popetimate, Birth Rate, Death Rate, International Migration Rate, Domestic Migration Rate, Net migration Rate)
- 131 Energy data (includes consumption, production, price etc for different types of energy). The different types of energy are - Biomass, Coal, Electricity, Fossil Fuels, Geothermal Energy, Hydropower, Natural Gas, LPG.
- 25 Economic Data (includes quarterly, yearly and average GDP)
- 6 Geographic data (State Code, Region, Division, Coast, Great Lakes)

The different rows consist of all the 50 states in the United States along with the data on the District of Columbia and the column totals data.

Variable transformation

1. Generating a year column

Many of our analyses required the year to be in a separate column. So we utilized the *Pivot_longer()* technique to extract the year from the column names and create it into one separate column. After which the *g_sub()* technique was used to remove the character values like “GDP” or “CoalC” from the column. This year column was then ready to be used for analysis. This was a way to clean our data and make it into a more usable format.

2. Year Difference

Almost all the variables in our dataset are marked with the year. For example, GDP2010-GDP2014. In our research process, we pay more attention to the difference between years than to a certain year. Thus, we have generated a lot of variables by subtracting the value of the previous year from the next year for example GDP_2010_2011(2011 GDP - 2010 GDP).

3. Year Percentage Difference

In our work, we didn't use this kind of transformation. In fact, at first, we were more inclined to year difference percentage ((next year-previous year)/previous year) because this method was more rigorous. But we later discovered that there were two problems with this approach. First of all, if we want to explore whether a 1000 population increase will bring about a 1000k GDP increase, year difference percentage is not very applicable here. Secondly, the year percentage difference will make some important values less obvious. For example, the GDP of Texas has increased the most, but its original GDP is also a lot, so in the year difference percentage, its value is very small.

4. Year Ratio

In the consumption analysis, the maps and distributions below, we used variables that represented year-over-year change. For instance, in the first year, 2010-2011, the values for 2011 were divided by the values from 2010 and multiplied by 100. This allows us to conclude that any values over 100.00 showed an increase from the previous year and any values less than 100.00 represented a decrease.

GDP Analysis

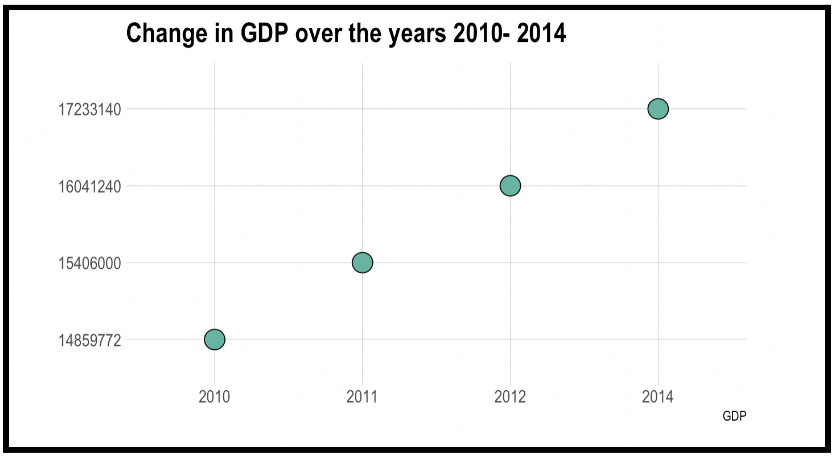
1. GDP Definition

Gross Domestic Product is the total market value of all goods or services produced in a market. We will be analyzing the citywise disparity in GDP and the possible causes/factors that lead to their GDP rise.

2. GDP Analysis

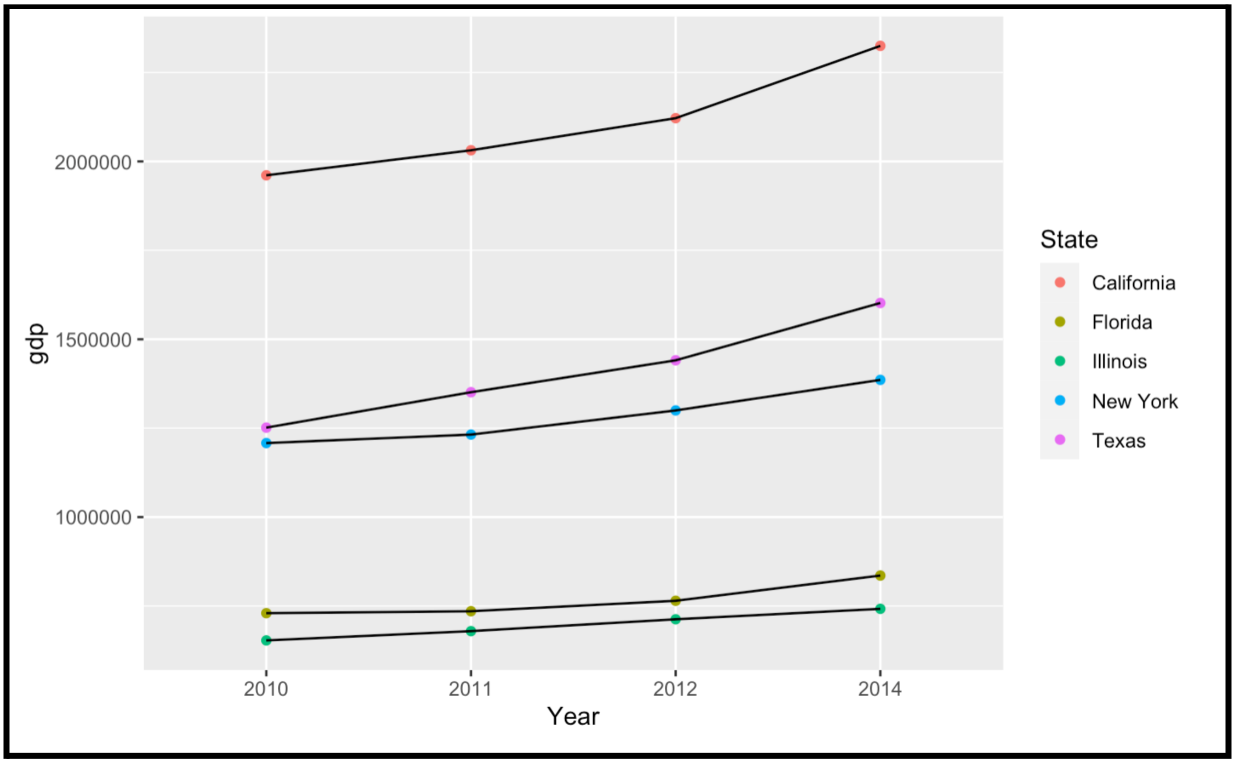
The data was analyzed to understand the trends in the United States GDP. We see that the GDP of the United States has risen by huge numbers over the years 2010-2014. This made us deeper into analyzing what caused this GDP rise and what were there any factors that significantly influenced the GDP rise.

The figure below shows the rise of GDP over the years:

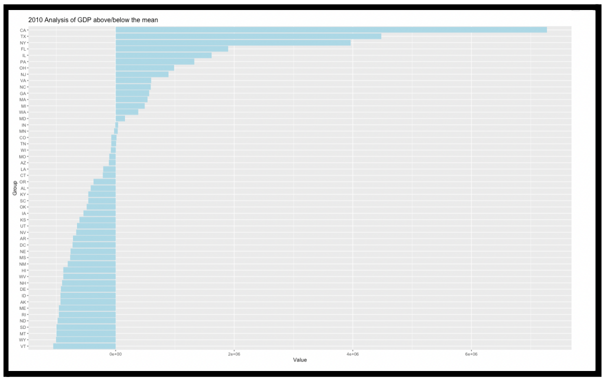


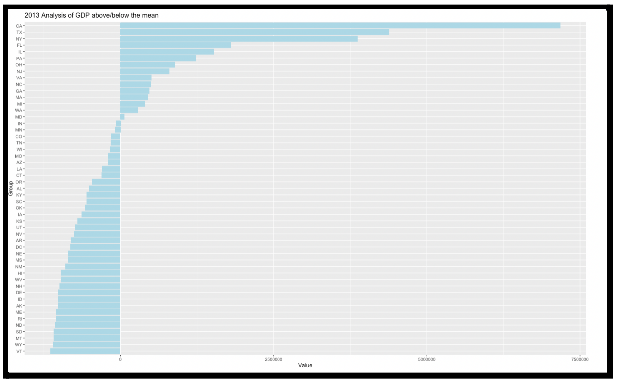
Next, the aim was to understand what were the cities with maximum GDP and plot the trends of their GDP increase/decrease. Through the years 2010 - 2014, California, Texas, New York, Florida and

Illinois have the highest GDP. The graph below shows how each of these cities changed in GDP value form 2010-2014:



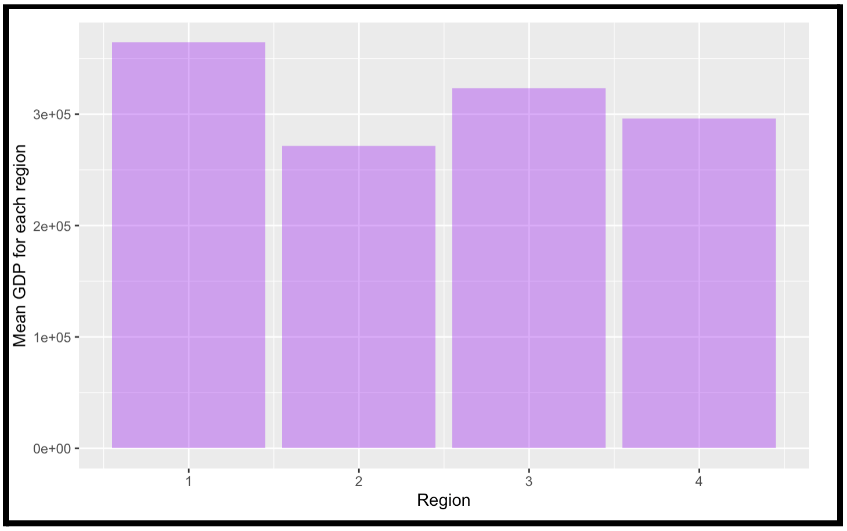
Another analysis we wanted to check for was which countries are below the mean GDP:





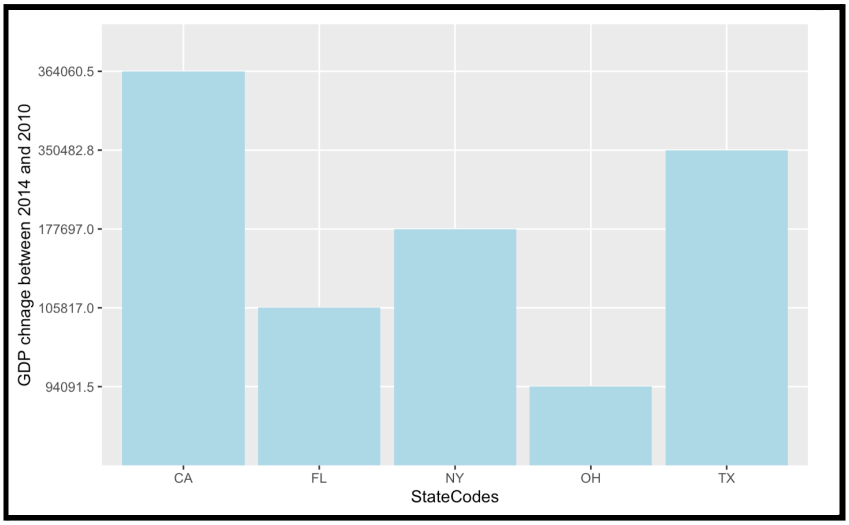
The bars protruding leftward depict states below the mean GDP and the states protruding rightward are states with high GDP(above the mean). The first graph is the analysis for 2010 and the second is for 2013. Vermont and Wyoming consistently have a GDP below the mean and California and Texas continue to maintain the lead. We see that the order of the states does not change, which means that the amount by which the GDP is increasing at an equal rate. Even though the mean per year is increasing, the city's GDP is also increasing. Hence now it is important to understand what is keeping these states to increase their GDP yearly? This question will be answered later in this paper.

Another important trend to notice was that the northeast part of the United States has significantly higher GDP than all other parts of the country. Second comes the Midwest part of the country, followed by south and west. The graph below emphasizes on these facts:



As stated above here we see the disparity in GDP for all 4 regions(1 = Northeast, 2 = Midwest, 3 = South, 4 = West).

The next analysis is to understand if there has been a huge jump in a city's GDP. The greatest difference in GDP for the years 2014 and 2010. The graph below shows that California, again, has the greatest GDP increase. Later in the report we will analyze these states, with the highest jump in GDP, and what caused this jump.

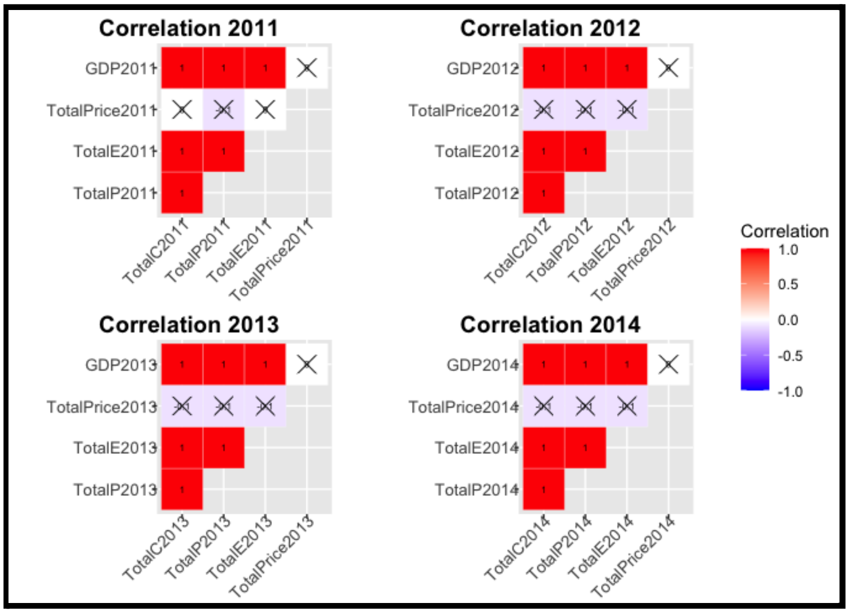


In this paper we will analyze what factors - consumption, energy or population causes an increase in the GDP. We will specially emphasize on the cities that we saw a huge rise in GDP over the years/ which have high mean GDP.

Energy Analysis

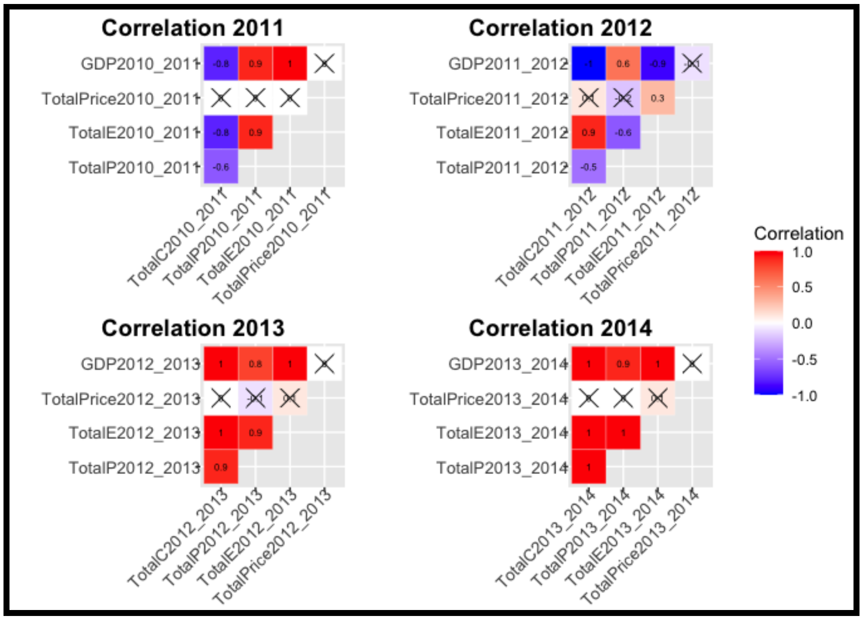
From the data, there are 131 variables under energy data, describing eight types of energy, including biomass, coal, electricity, fossil fuel, geothermal, and hydropower, from four perspectives, which are consumption, production, expenditure, and price change. We wonder how these four perspectives relate to the GDP and which entries are the main drivers for each perspective.

As the data include years of information, we focus on yearly exploration rather than cumulative analysis. We wonder if consistent correlation existed from 2010 to 2014 so that we could use such analysis for future GDP performance prediction. Here, we consider the correlation between the yearly GDP and year energy group amount. From the plot, we can find the GDP has a strong positive relationship with consumption, production, and expenditure for energy. However, there is no relationship between GDP and average price.



As we know three groups here drive GDP, we wonder if yearly change of each group had a similar relationship with GDP change. Thus, the following correlation plot focuses on change. For each energy yearly change, we find the difference between two years. From the plot, we can find that GDP has a strong positive relationship with energy production. Relationships between GDP and consumption were strongly negative in 2011, 2013, and 2014, but changed to strong positive in 2012. Relationships between GDP and consumption were strongly negative in 2011 and 2012, but changed to strong positive in 2013 and 2014. However, there is no relationship between GDP change and average price change.

Thus, we believe states with more energy consumptions, productions, and expenditures may have more GDP. states with large energy production will have more GDP Change. However the correlation may be impacted by specific energy or some outliers. Thus even if we can find a strong relationship, some groups may have used misleading information. For the next part, we are going to discuss each perspective.



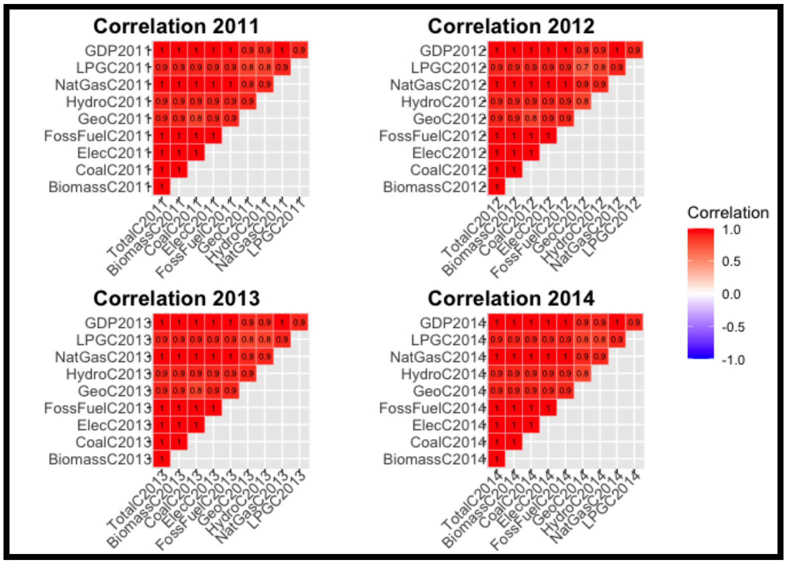
1. Consumption:

Consumption is the first category describing the amount of consumption in each given year. From the data, the US uses all eight energies across the country. Before we look deep about specific energy distribution, we wonder how each consumption relates to the Year GDP.

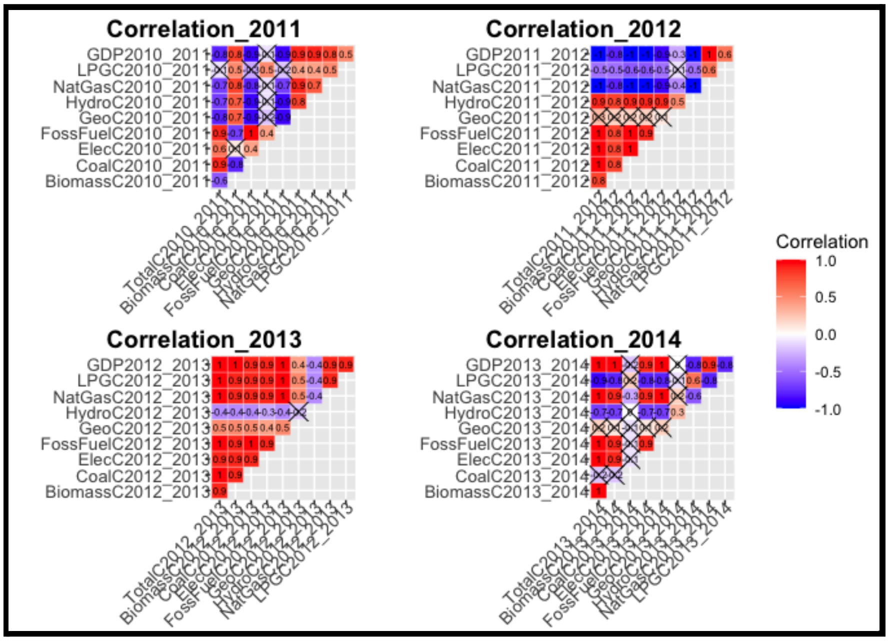
1.1 Correlation Analysis:

For each year, we have coal consumption, biomass consumption, electricity consumption, fossil fuel consumption, geothermal consumption, hydropower consumption, and total consumption. Here we consider two correlation analysis, consumption changing analysis and yearly consumption analysis.

The first is yearly consumption analysis. Here we only consider yearly GDP, and each energy consumption in the year. We can find that consumption of all energy consumption has a strong positive relationship with GDP, indicating higher energy consumption will have a higher GDP.

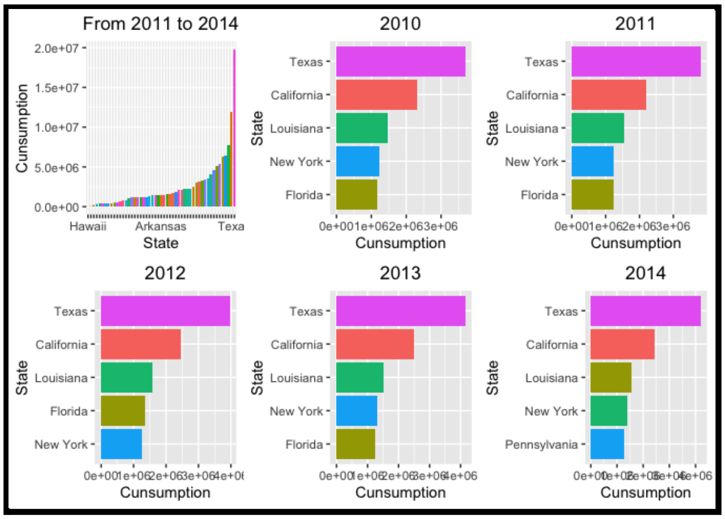


The second is yearly consumption changing analysis. To calculate the GDP in the given year, we find the difference between total GDP in the given year and total GDP in the last given year. We apply the same strategy to find each energy consumption change in the given yearThen, we are able to find the correlation rate and p-value for selected data. Here are the plots containing all four year correlations.



From the plots there is a strong positive correlation between GDP and consumption change in 2013 and 2014, but has a strong negative correlation in 2011 and 2014. In 2013 and 2014 more energy types had positive correlation as total energy consumption had. However, in 2011 and 2012 more energy types had negative correlations than energy types have positive correlations here.While each energy consumption change performs a different correlation with GDP. We find that GDP always has a strong positive correlation with natural gas. Here we are going to select Natural gas consumption as a representative for the three major drives, exploring its distribution.

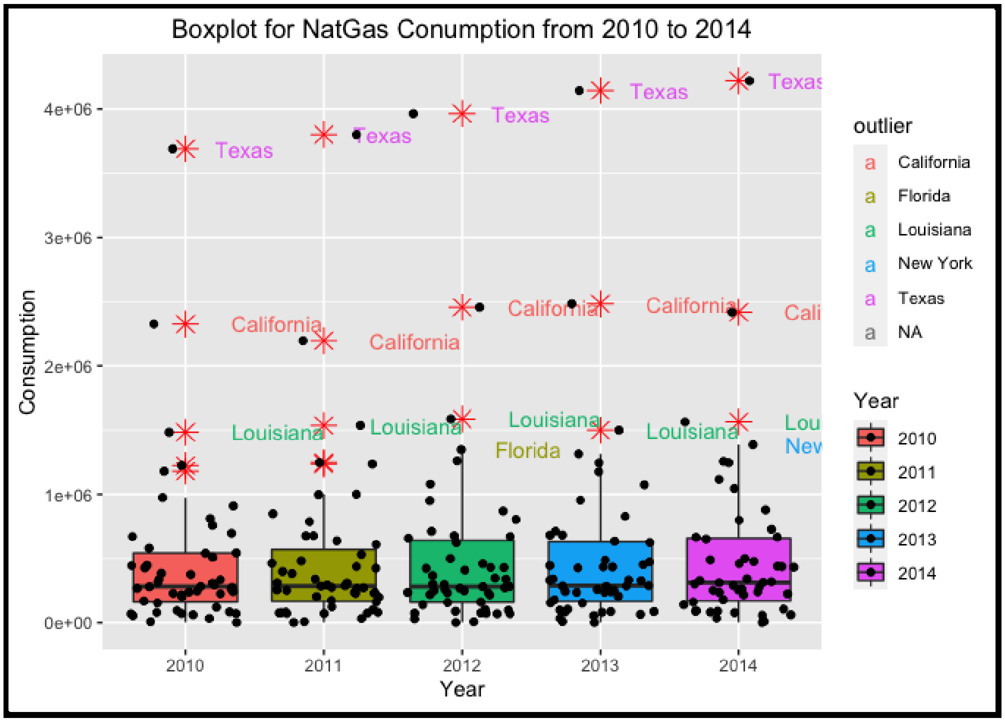
2. Natural Gas Consumption Histogram Analysis



The first step here is to better understand how Natural Gas is distributed nationwide. From 2010 to 2011, Texas consumed the most natural gas, leading California every year. While the top three states, Texas, California, and Lousina, lead other states, New York and Florida consume less energy. In 2014, Pennsylvania replaced Florida, taking the position at the top 5 consumption states. The next step is to explore more statistical information.

2.1 Natural Gas Consumption Boxplot Analysis

The Following box plot gave us a direct summary to identify mean of natural gas consumption in the given year. We can see that the mean consumption does not change too much. To find the outlier states, we set a function to find outliers if the state is not in the 50% confidence interval.



2.2 Consumption Analysis conclusion

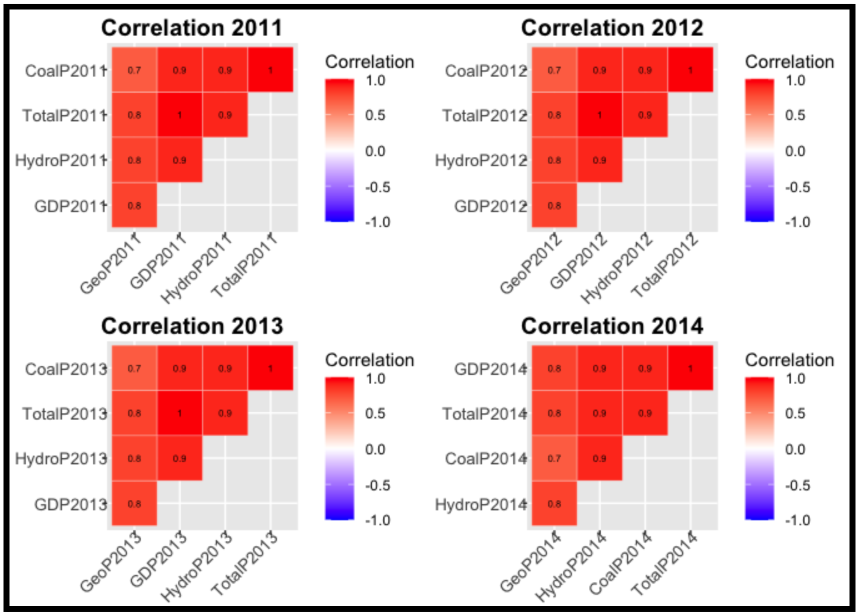
From the plots previous, we find there is a strong positive relationship between GDP and each energy consumption. However, each energy consumption yearly change has a varied correlation with yearly GDP change in the given year. Each energy consumption change would impact the overall correlation with GDP change. However, no matter positive or negative, the relationships between all consumption changes and GDP change are always strong. For natural gas consumption change, it always has a strong positive relationship with GDP change, meaning producing more consumption will have a higher GDP. In the given year, Texas consumed the most natural gas and kept its position all the time. Based on the correlation GDP has a positive relationship with energy consumption, the outliers here should have a higher GDP than other states. However, it could not match the GDP rank, we may still need to look at other drivers.

3. Production

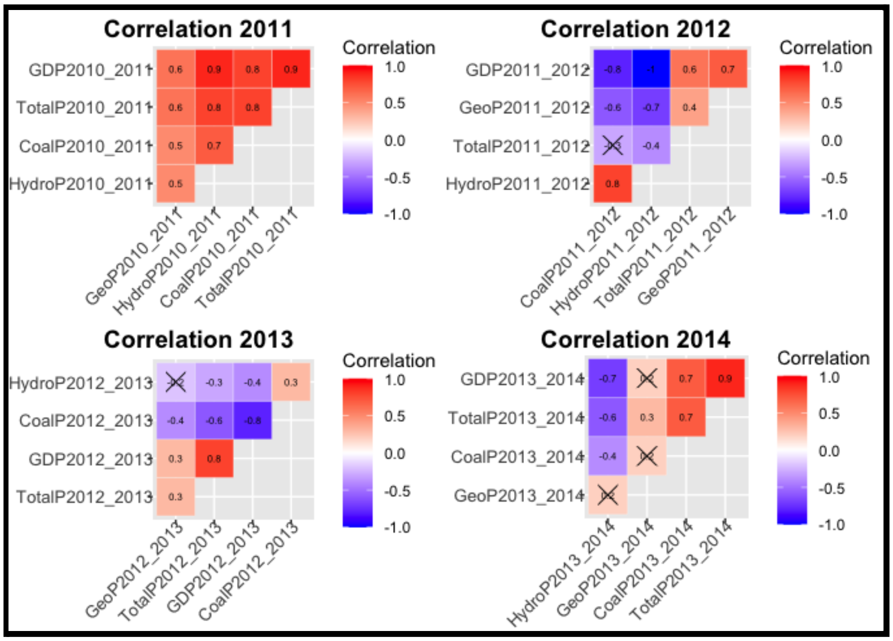
Production is another big topic describing the amount of production in each given year. Different from consumption data, not all eight energies are produced in the US in the given year. Literally, from the data, only geothermal energy, hydropower energy, and coal energy are produced in the US. Similar to our consumption analysis, the first step is to consider if there is a relationship between GDP and typical energy production in the given year and a relationship between GDP change and typical energy production change in the given year.

3.1 Correlation Analysis

The first step here's to find each energy production and GDP in the given year. We can find that all energy productions have high relationships with the given year GDP.

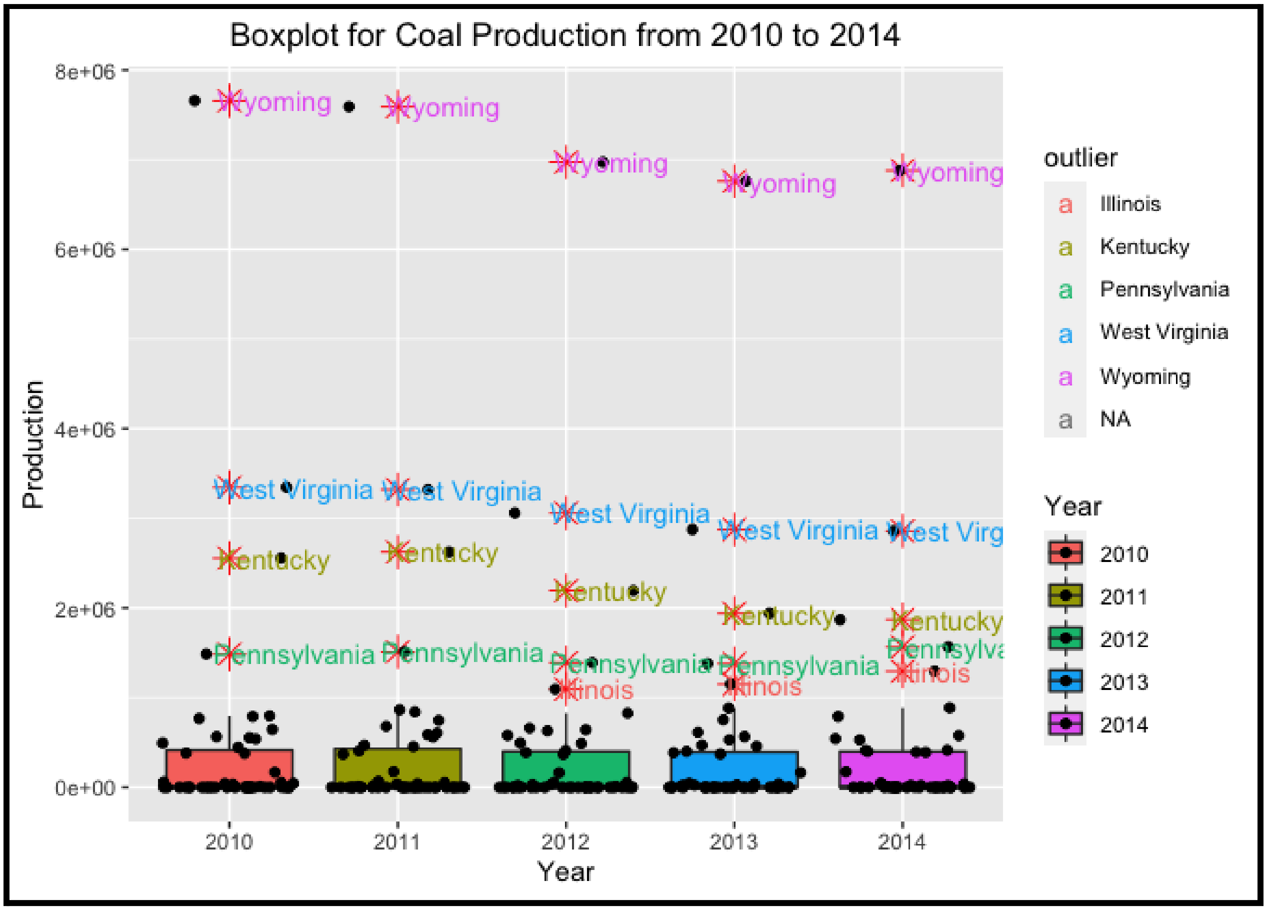


The second step here is to find each energy yearly production change and GDP change. From the following group, we summarise that there is a strong positive correlation between GDP change and production change in 2011, 2012,and 2014, but a strong negative relationship in 2013. yearly GDP has a similar relationship with coal energy.Thus, we are going to use coal consumption as a representative. Then the next step is to explore more details about coal energy production distribution, analyzing if the outliers will impact the hyposis.



Coal Energy Production Box Plot Analysis

While we find there is a strong relationship between GDP and coal energy production, we find that the mean of state production is 25000 here, which did not change a lot in the given years. We can find the outlier here is Texas, Indian, and ohio. Texas has the most production and far more than other states have.



3.2 Production Analysis conclusion

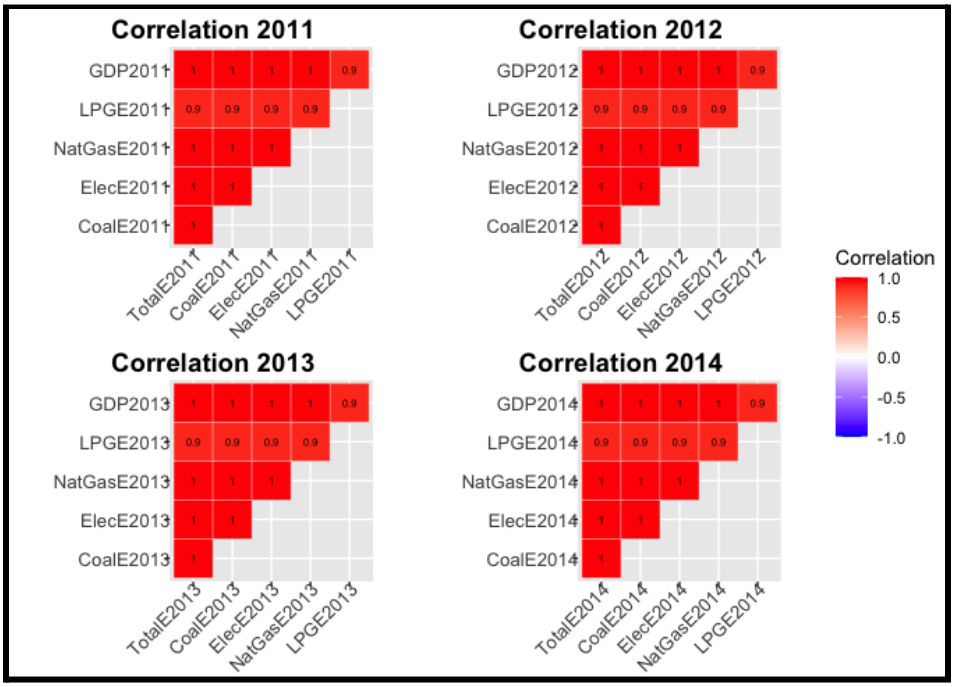
While The relationship between production and GDP changes yearly, we find there is a consistent positive correlation between GDP change and coal production change. We doubt this prediction. From the box plot, we find the leading state is Wyoming. However, when we look at the mean, we find most states do not produce any coal. It implies that the correlation result will be impacted by few states. If we remove the outlier here, we could find GDP has nearly no relationship with energy production, as the mean of production is zero. When we want to consider using energy consumption to predict GDP, we will not use energy production as a factor.

4. Expenditure

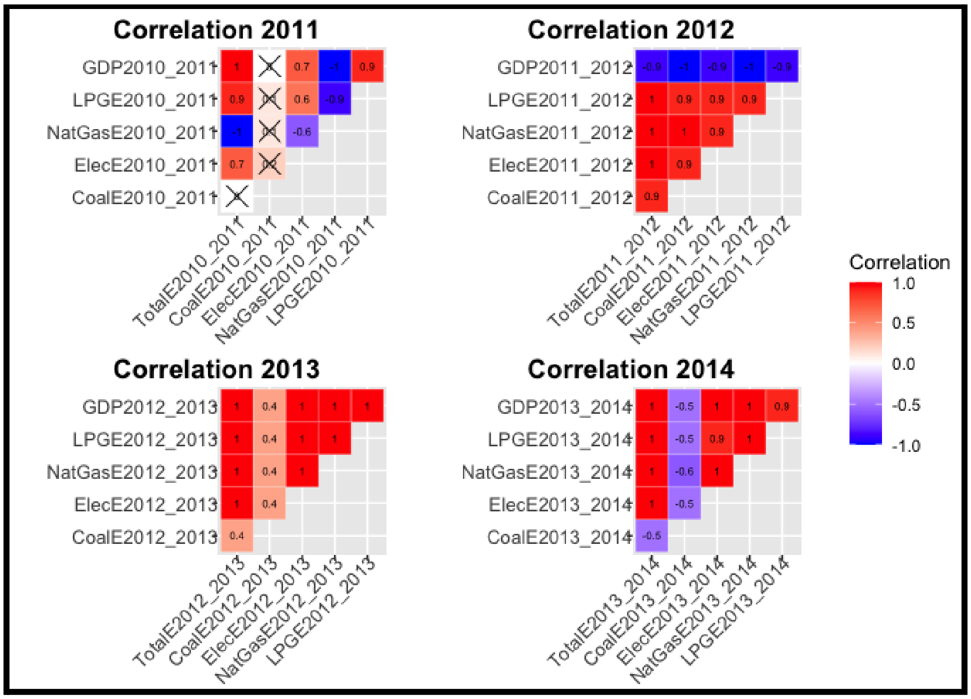
Expenditure will be the next topic describing the amount of expenditure in each given year. Similar to energy production, we do not have expenditure data for all energies. Here, we find LPG expenditure, natural gas expenditure, electricity expenditure, and coal expenditure in the given year.

4.1 Correlation Analysis

The first step here is to find each energy production and GDP in the given year. We can find that all energy productions have high relationships with the given year GDP.

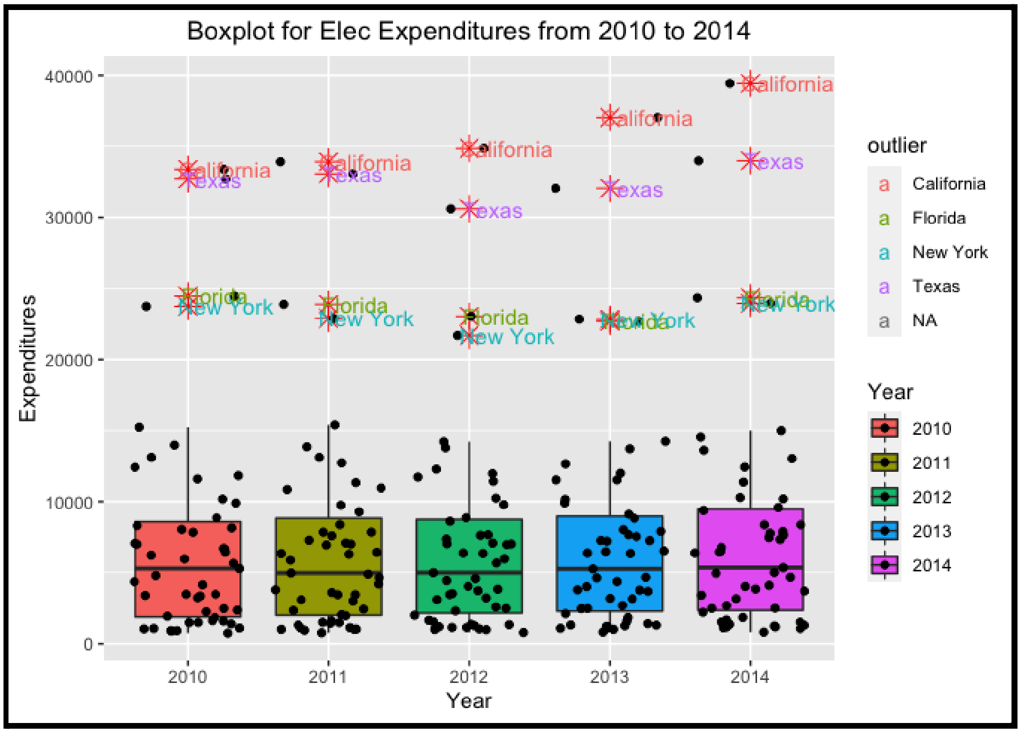


The second step here is to find each energy production change and GDP change in the given year. There is a strong positive correlation between GDP and expenditure except 2012. The solution is close to our first correlation analysis just made. While all four energy expenditures show a positive relationship with the given year GDP, expenditures in electricity and LPG keep extremely strong positive correlations with GDP in all four years Thus, we want to select electricity as representative for a deep look.



4.2 Electricity Expenditure Box Plot Analysis

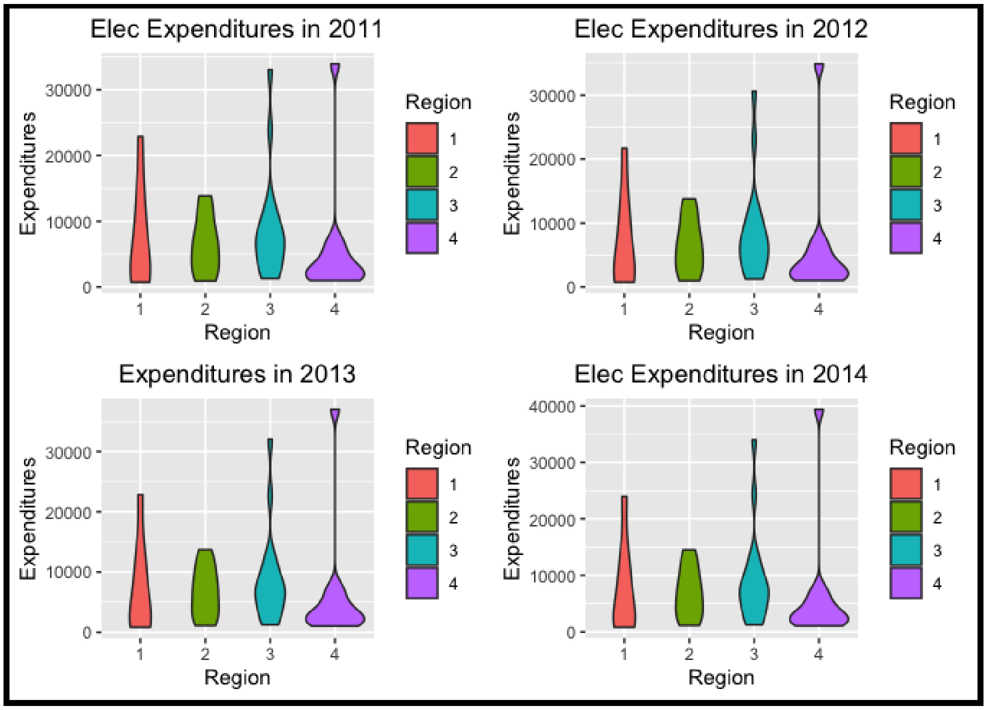
The following plot shows the mean and outlier of electricity expenditure across the US based on the given year. With our outlier function, we could find California, Texas, Florida, and New York have had the most expenditures in the given year. With the correlation we had, We me assume that these four states have more GDP than the remaining states.



As we know, the dataset separates states into four regions, 1 = Northeast, 2 = Midwest, 3 = South, 4 = West. Both Florida and Taxes are in the south, New York is in the Northeast, and California is in the West. We wonder about the statistics information from the region for more inspiration.

4.3 Regional Electricity expenditure violin plot based on year

Different from the boxplot for the whole nation analysis, regional violin plots in the given year focus on expenditure distribution in the given year. From the plots, we could find 1- Northeast and 2- Midwest have a more concentrated distribution among its cities. Distributions in 3-South, and 4- West are impacted from outliers. For 3- South, the outliers will be Florida and Texas. For 4-West, the outlier is California.

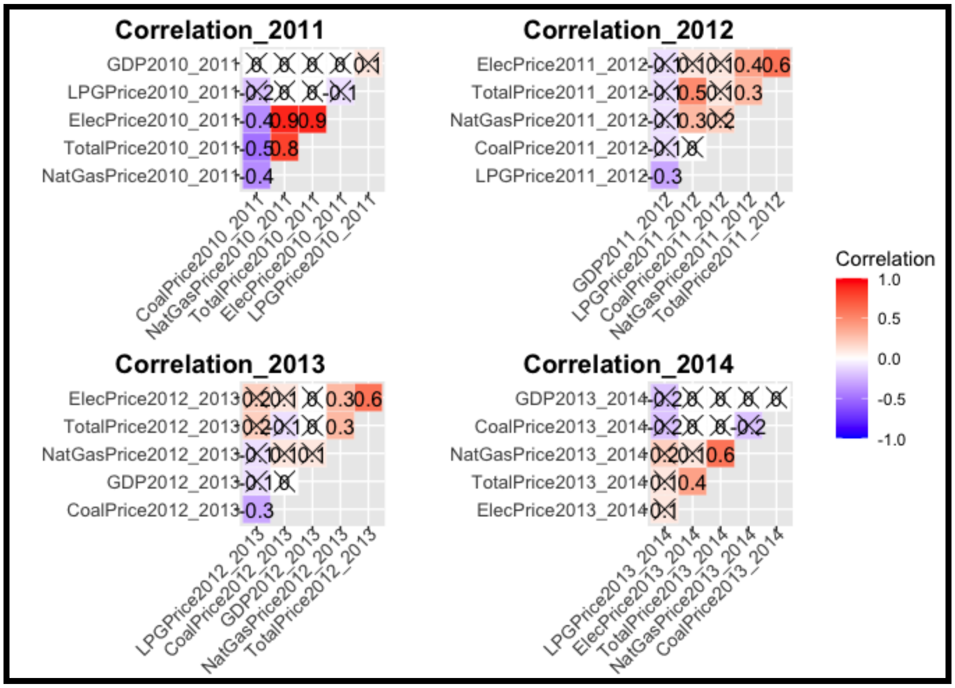


4.4 Expenditure Analysis conclusion

As a representative, we can find that electricity expenditure has a positive relationship with GDP performance. More expenditures mean higher GDP performance. California, Texas, Florida, and New York have had the most electricity expenditures in the given year. As we find there is a positive relationship between expenditure and GDP, the pot indicates such four states will have a higher GDP performance. These states are outliers of their own regions, indicating they have higher GDP performance in their own regions.To consider the relationship between expenditure and GDP, we believe each energy type has a positive relationship. Higher expenditure will indicate more GDP. The ranking result is very close to the real GDP rank. We may consider it is a important factor determining GDP.

5. Price Change

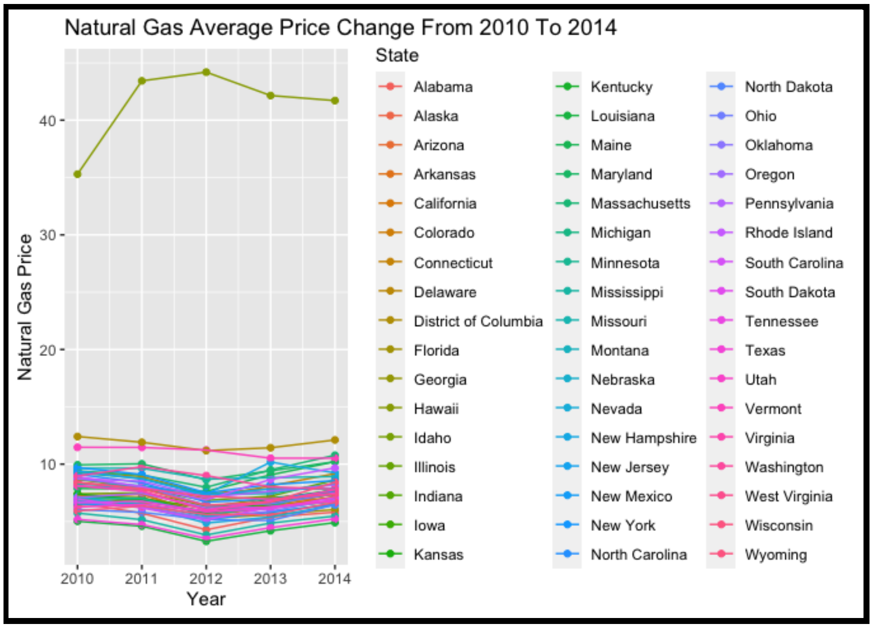
Price is the last group in energy describing the average price performance in each given year. While we believe each energy price should be available,we do not have price data for all energies. Here, we have coal price, electricity price, and natural gas price, and total average price in the given year. However, directly looking for correlation between price and GDP is meaningless as yearly price is a reflection of price change. Thus we want to find the price difference in the given year and compare it with yearly GDP. To find the yearly price change, we find the price difference between two years.



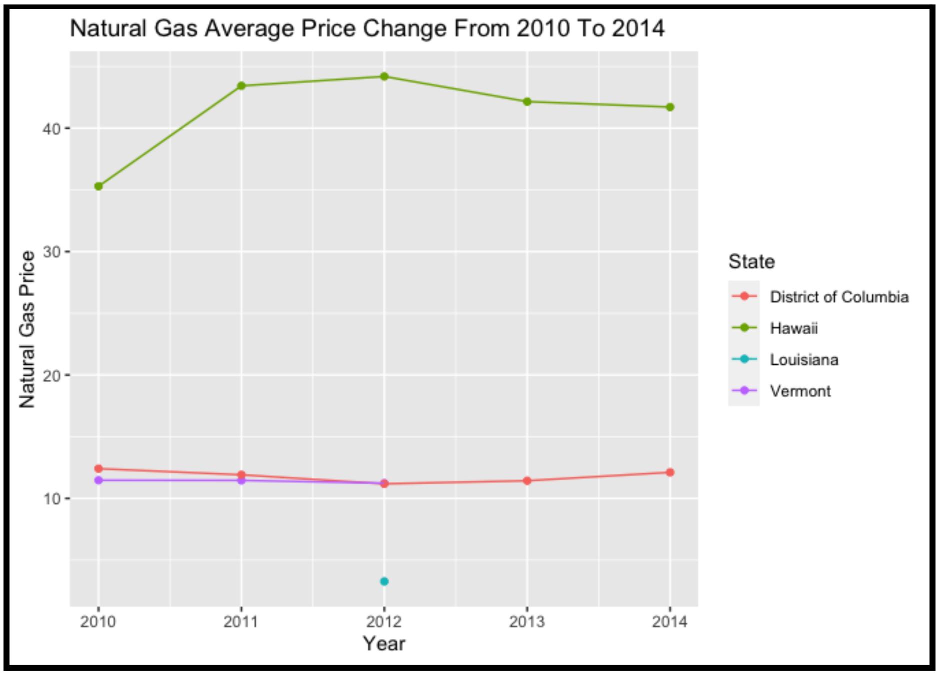
From the correlation plots, there is no correlation between GDP and price change in 2011 and 2014, but there are positive correlations in 2013 and 2014. We find the GDP does not have a relationship with any energy price change in 2011. GDP has a positive relationship with nat gas price in 2012 and 2013. However, such a relationship faded in 2014. Thus, we can assume that natural gas price will be the driver impacting total price change in the same year, impacting the GDP.

5.1 Natural Gas Price Time Series Analysis

The first step here is to visualize how each state's price changes across the five years. For most states, there is a price decrease in 2012 and a price increase later. However, We can find that one state had an inverse performance, increasing its natural gas price from 2010 to 2012 and decreasing price from 2012 to 2014. Additionally, while we find there are price changes across the US in the given year, the general tendency is inelastic. However the outlier could bring significant change.



We will use the outlier function again, to find the outlier whose confidence interval is 50% here. We find Hawaii and the District of Columbia are outliers for each year. Louisiana is an outlier in 2012 and Vermont is an outlier from 2010 to 2012. From the time series table, we find Hawaii's natural gas price impact overall price change tendency. The outlier can bring a basis correlation between price change and GDP.

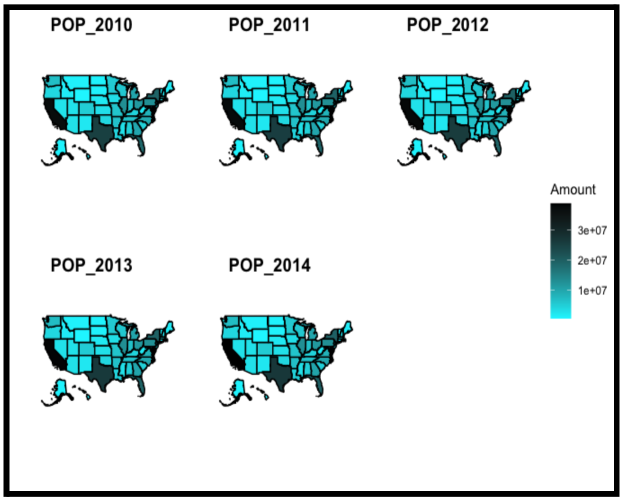


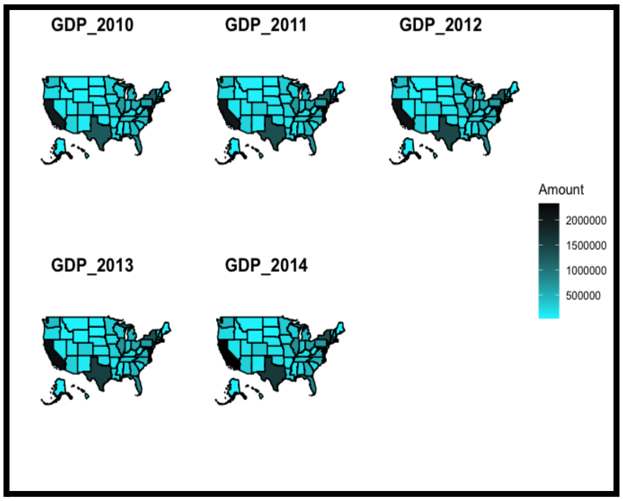
5.2 Price Change Conclusion

Generally, we could not find correlation between price and GDP. While some states may set different price changes, the outlier impact indicates a misleading correlation in the 2012 and 2013 period. Such impact should be prevented from conclusion.

Population Analysis

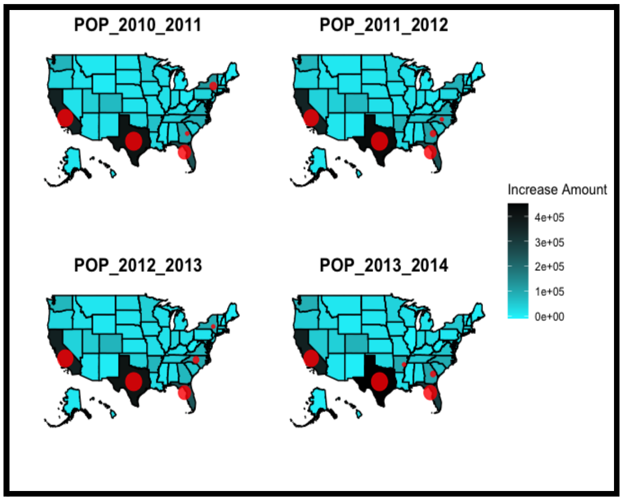
Firstly, we want to know, in general, whether the population has a correlation with GDP. In this case, we use maps to show the population and GDP situation from 2010 - 2014.

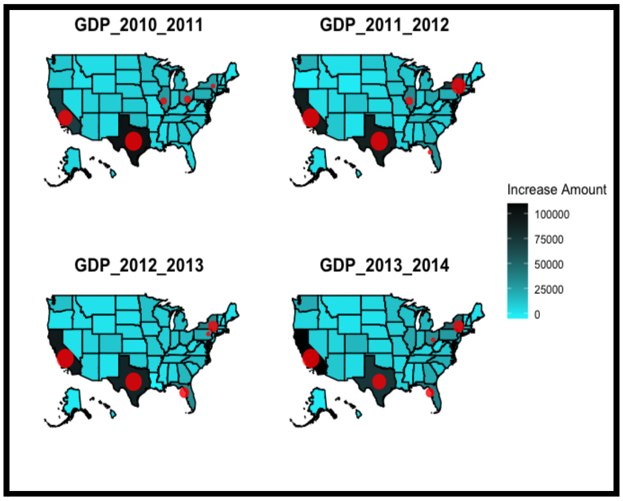




Based on the above graphics, we can see two things. The first thing is that the population and GDP didn't change too much in these five years. The second thing is that the population map is very compatible with the GDP map. Thus, we can conclude that population has a very strong correlation with GDP, which makes sense because it's reasonable that more people will generate more GDP.

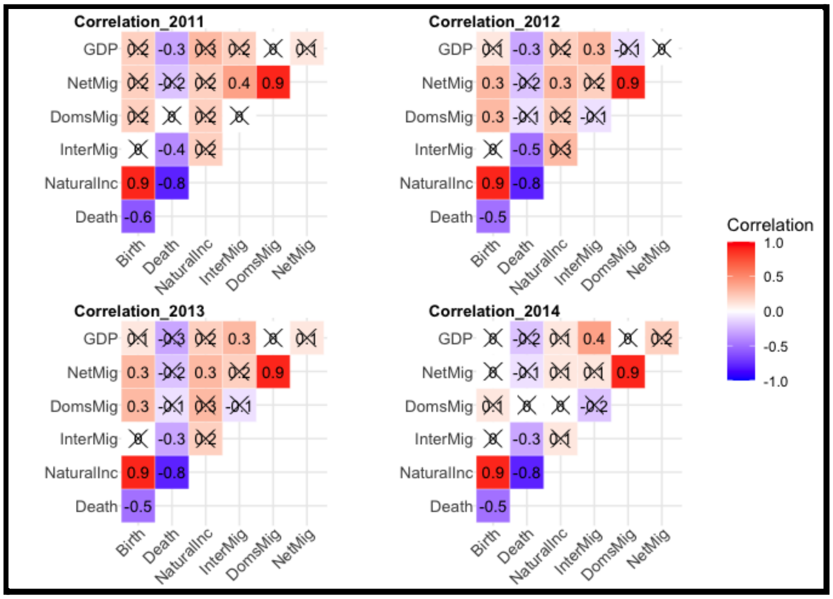
Secondly, we want to further understand the relationship between population and GDP by studying whether the population increase will contribute to the GDP increase.





We use red circles to mark the top 5 states with the largest population increase and GDP increase. There are three things we can find from the above graphics. The first thing is that from 2010 to 2014, the population and GDP of all states have increased. The second thing is about the overlaps in top 5 states from population and GDP. The first map has 3 overlaps, the second map has 3 overlaps, the third map has 4 overlaps, and the fourth map has 3 overlaps. The overlap percentage is relatively high. The third thing is that the two states with the largest population increase also have the largest GDP increase. From the above three things, we can conclude that the population increase is likely to contribute to the GDP increase and the amount of population increase will positively influence the amount of GDP increase. When population increase is significant enough, this influence will become greater.

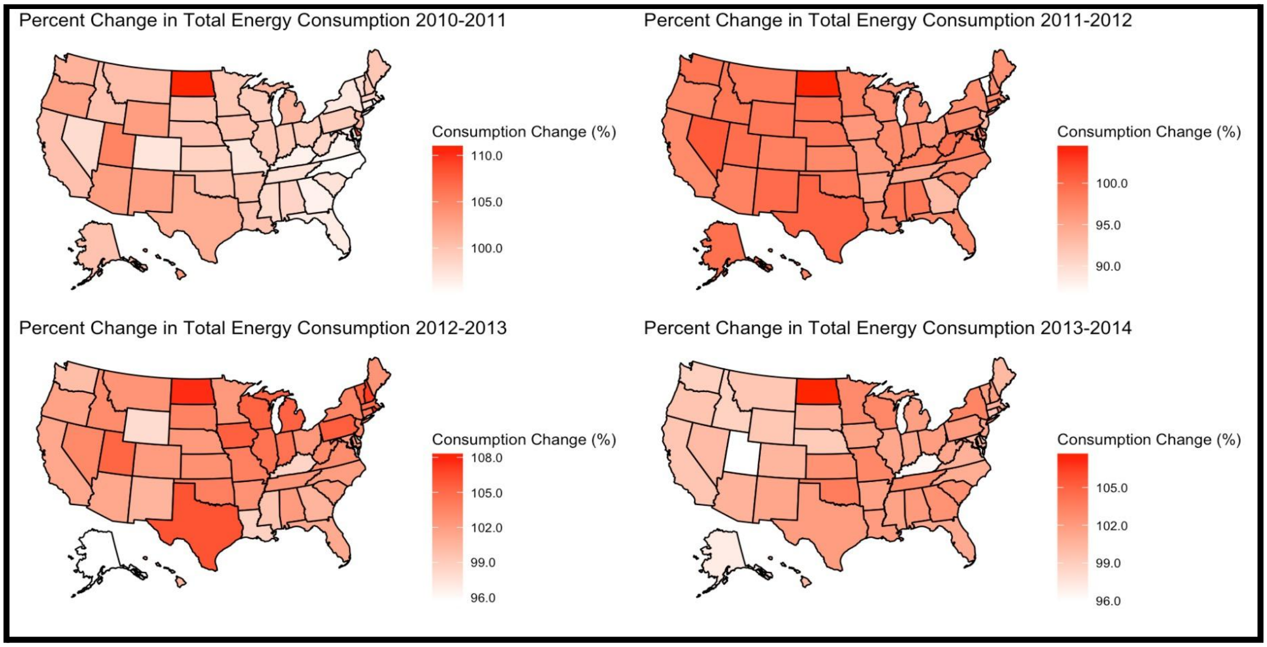
Lastly, we want to focus on more details. We want to know which factors that affect the population have a stronger correlation with GDP. These factors are birth rate, death rate, natural increase rate, net international migration rate, net domestic migration rate, net migration rate.



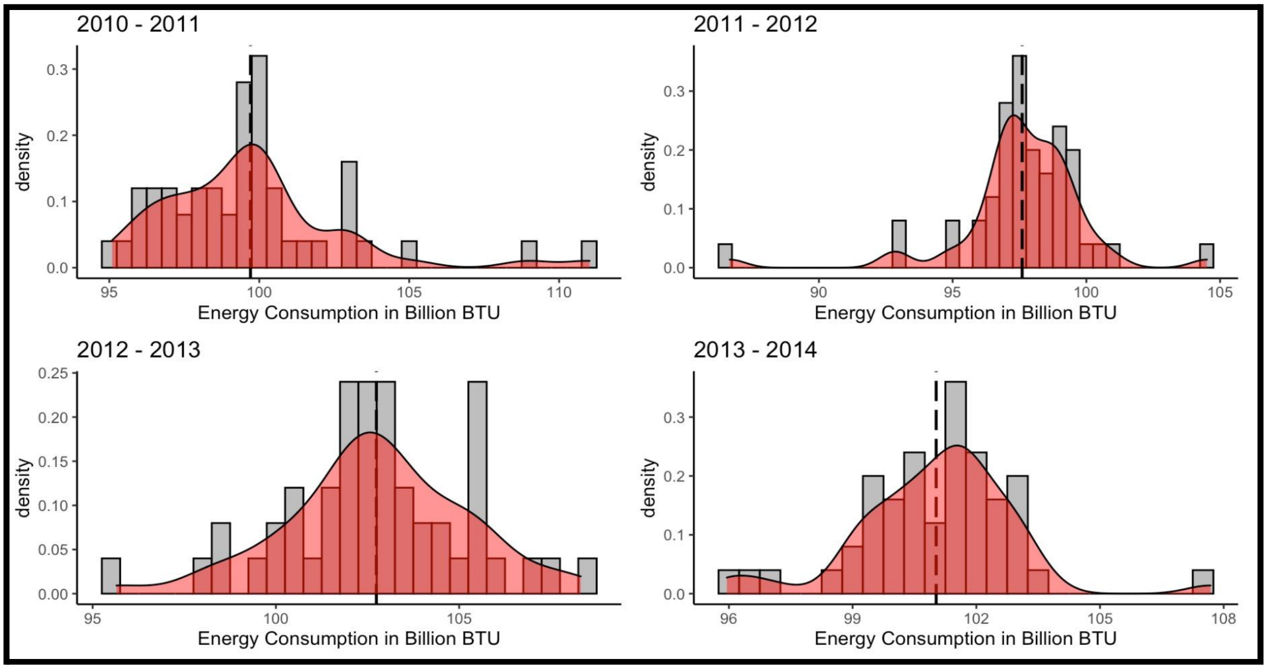
No significant correlations are barred based on the correlation p-values. In 2011, the left factor is death rate; in 2012, the left factor is death rate and net international migration rate; in 2013, the left factor is net international migration rate; and in 2014, the left factor is net international migration rate again. Thus, we can conclude that the death rate and net international migration rate have stronger correlations with GDP. The death rate negatively influences the GDP and the net international migration rate positively influences the GDP.

Detailed Consumption Analysis

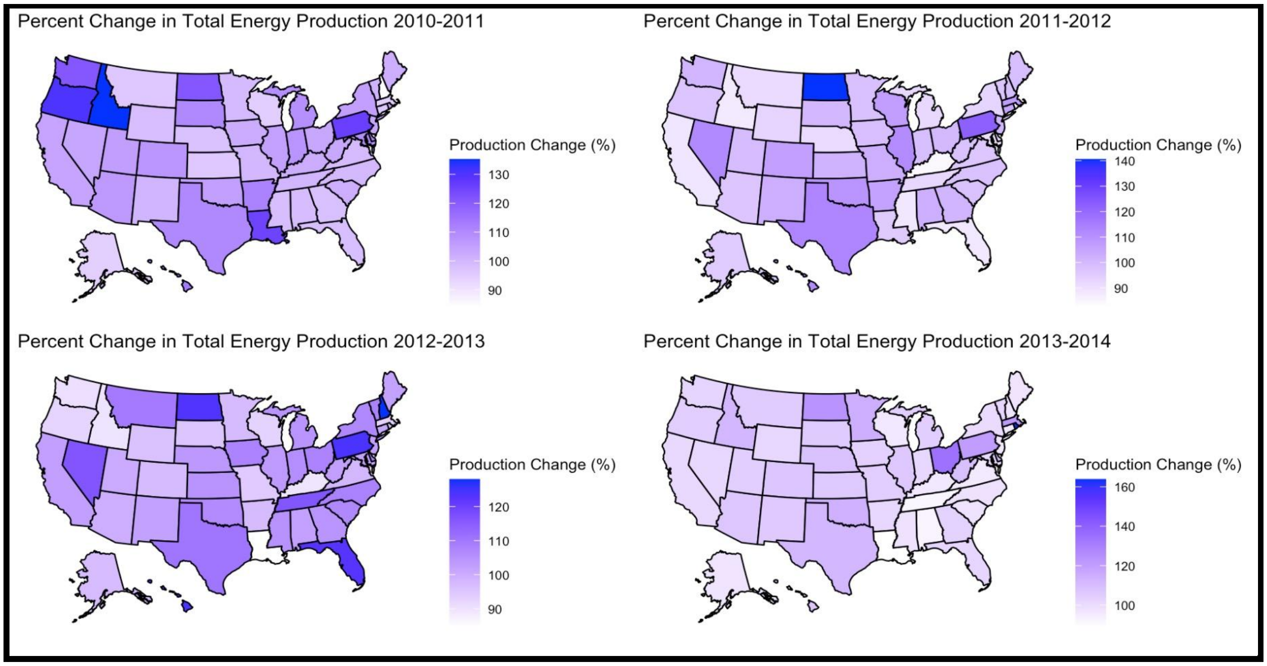
Individual State Year-Over-Year Changes

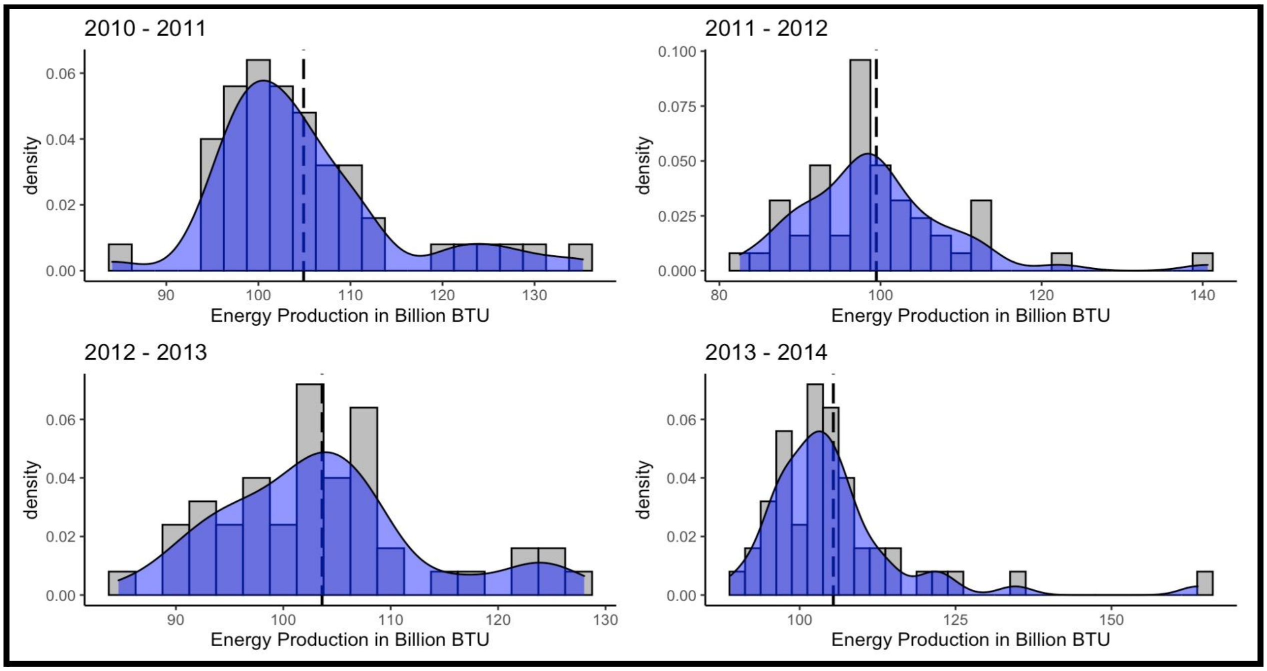


Percent Change in Total Energy Consumption 2010-2011
Consumption Change (%)
110.0
105.0
100.0
Percent Change in Total Energy Consumption 2011-2012
Consumption Change (%)
100.0
95.0
90.0
Percent Change in Total Energy Consumption 2012-2013
Consumption Change (%)
108.0
105.0
102.0
99.0
96.0
Percent Change in Total Energy Consumption 2013-2014
Consumption Change (%)
105.0
102.0
99.0
96.0

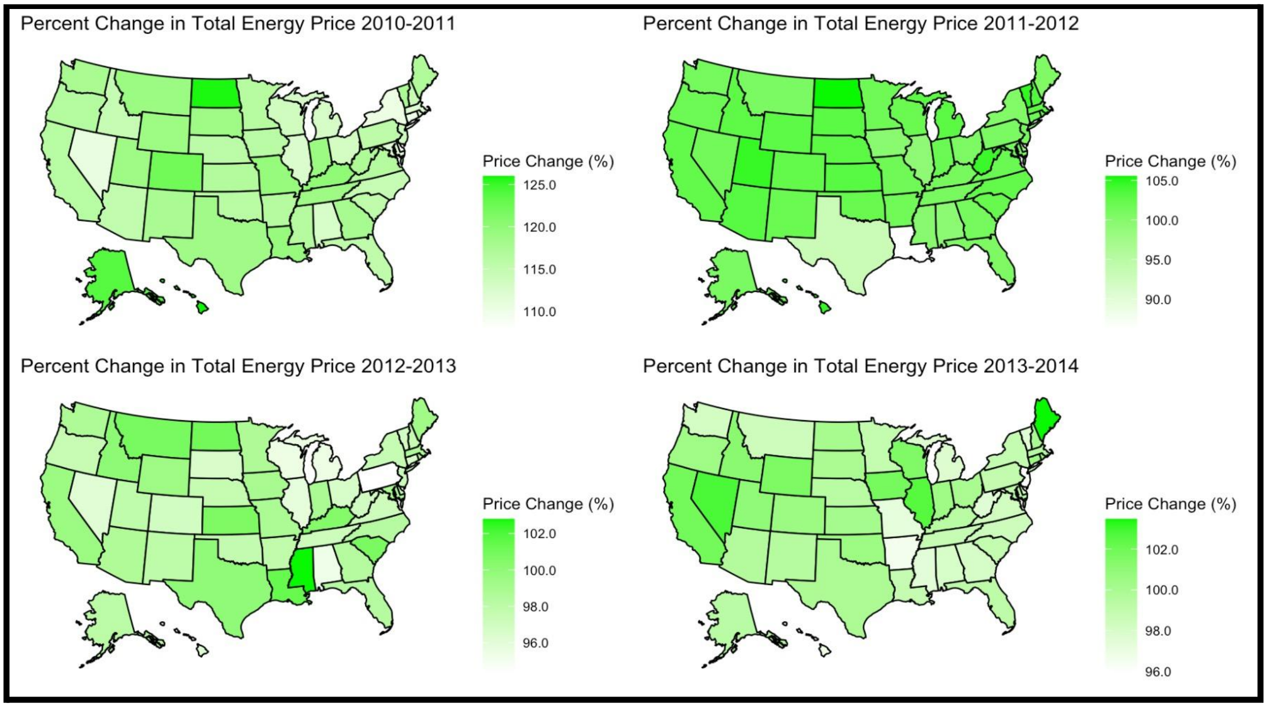


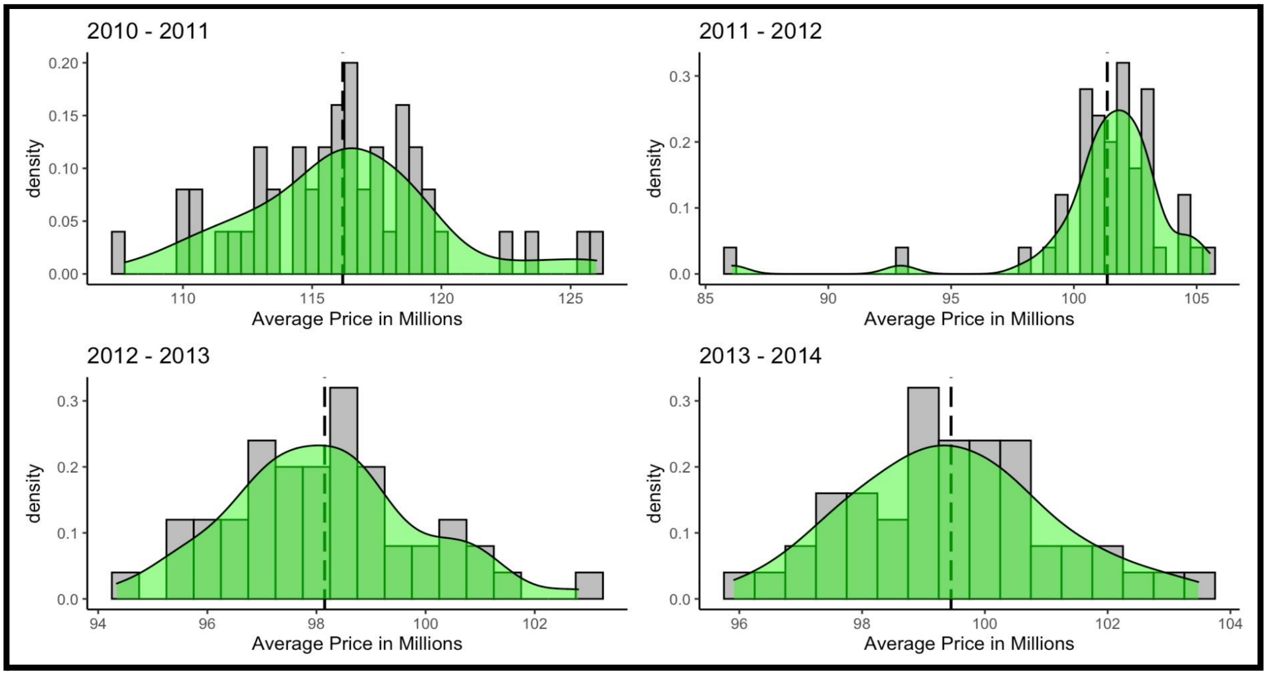
In the above graphs, we see the percent change (next year / previous year * 100 %) in total energy consumption year over year from 2010 to 2014. The most noticeable state from the maps is North Dakota. While North Dakota ranks 48th in the United States in population, it consistently leads the country with the highest percent change in total energy consumption year over year. In the graphs above, we can see how each given year's percentage changes are distributed. The mean percentage change of the total energy consumption for each year is represented by a black dashed line. 2012 is the only year represented where the mean energy consumption by state was less than the year before.


Percent Change in Total Energy Production 2010-2011
Production Change (%)
130
120
110
100
90
Percent Change in Total Energy Production 2011-2012
Production Change (%)
140
130
120
110
100
90
Percent Change in Total Energy Production 2012-2013
Production Change (%)
120
110
100
90
Percent Change in Total Energy Production 2013-2014
Production Change (%)
160
140
120
100

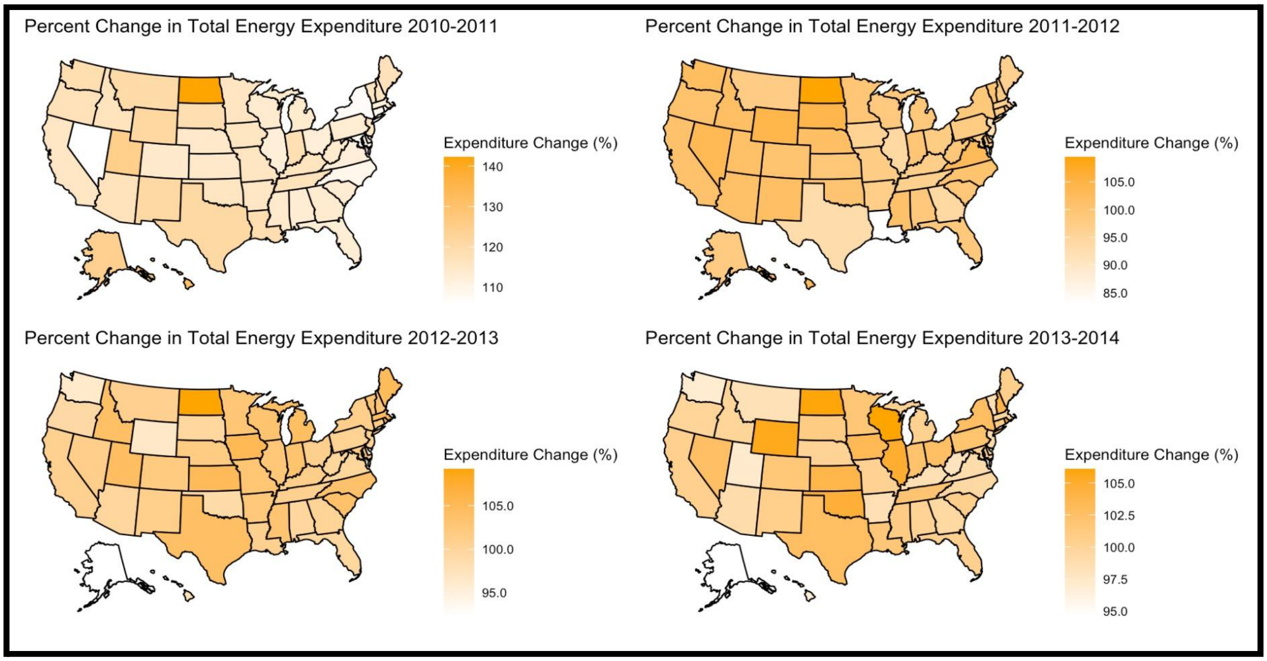


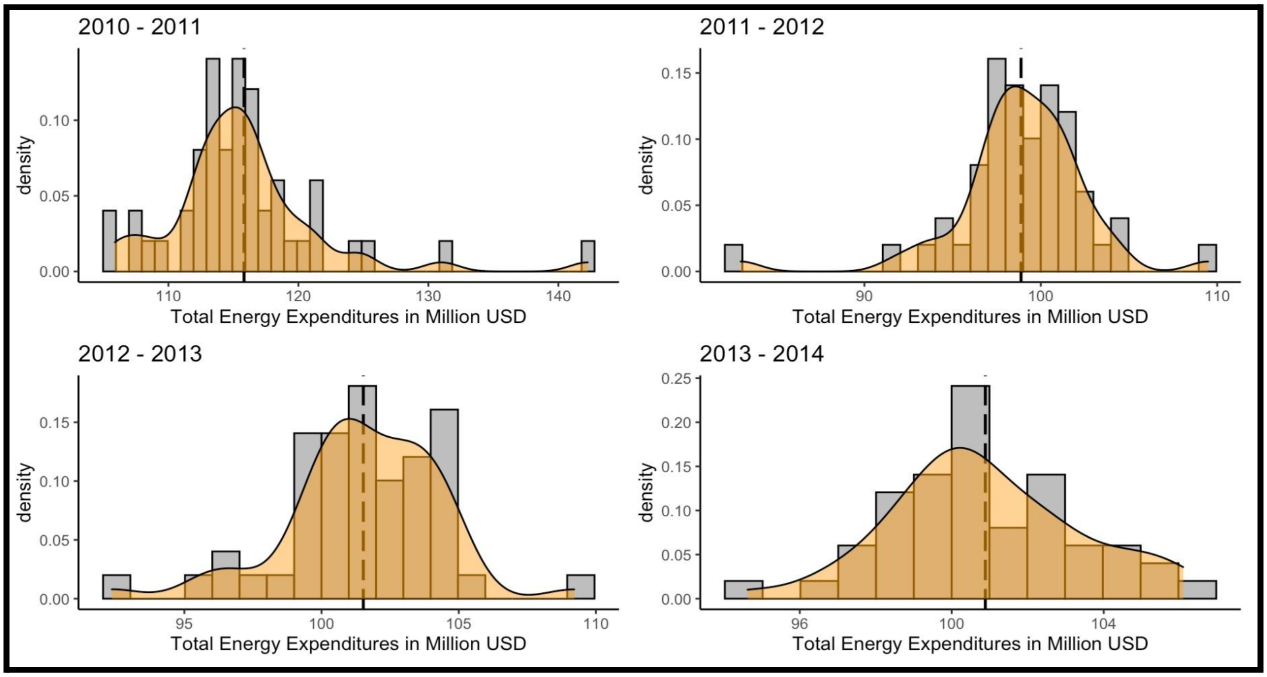
In the above graphs, we see the percent change in total energy production year over year from 2010 to 2014. Consistent with the graph above, we noticed that North Dakota is noticeable in this graph. We see a deep blue in two of the graphs, which signifies one of the highest changes, year over year, in total energy production. Also worth noting is how Oregon and Louisiana had a huge increase from 2010 to 2011, followed by a massive drop to near the bottom of the United States between 2011 and 2012. The mean percentage change of the total energy production for each year is represented by a black dashed line. Each distribution shows us that the mean energy production increased in years 2011, 2013, and 2014, while staying relatively the same in 2012.


Percent Change in Total Energy Price 2010-2011
Price Change (%)
125.0
120.0
115.0
110.0
Percent Change in Total Energy Price 2011-2012
Price Change (%)
105.0
100.0
95.0
90.0
Percent Change in Total Energy Price 2012-2013
Price Change (%)
102.0
100.0
98.0
96.0
Percent Change in Total Energy Price 2013-2014
Price Change (%)
102.0
100.0
98.0
96.0



In the above graphs, we see the percent change in total energy price year over year from 2010 to 2014. While price does not have a significant impact on gross domestic product, we wanted to have it represented here to show how Americans dealt with increasing prices from the fallout of the recession of 2008. We noticed a significant increase in prices across the board, with only a few exceptions, for 2011 and 2012 and concluded this to be a result of the recession. The mean percentage change of the total energy price for each year is represented by a black dashed line. The biggest increase in mean production happened in 2012, while we saw a slight decrease in 2013 and 2014.

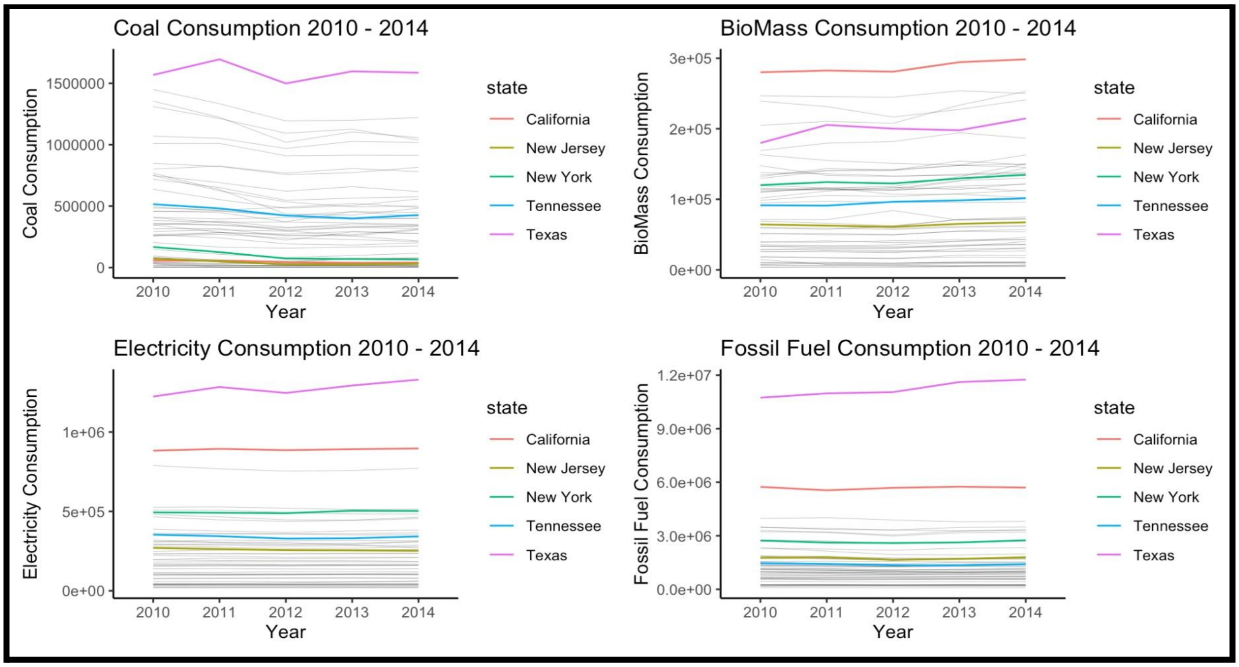

Percent Change in Total Energy Expenditure 2010-2011
Expenditure Change (%)
140
130
120
110
Percent Change in Total Energy Expenditure 2011-2012
Expenditure Change (%)
105.0
100.0
95.0
90.0
85.0
Percent Change in Total Energy Expenditure 2012-2013
Expenditure Change (%)
105.0
100.0
95.0
Percent Change in Total Energy Expenditure 2013-2014
Expenditure Change (%)
105.0
102.5
100.0
97.5
95.0

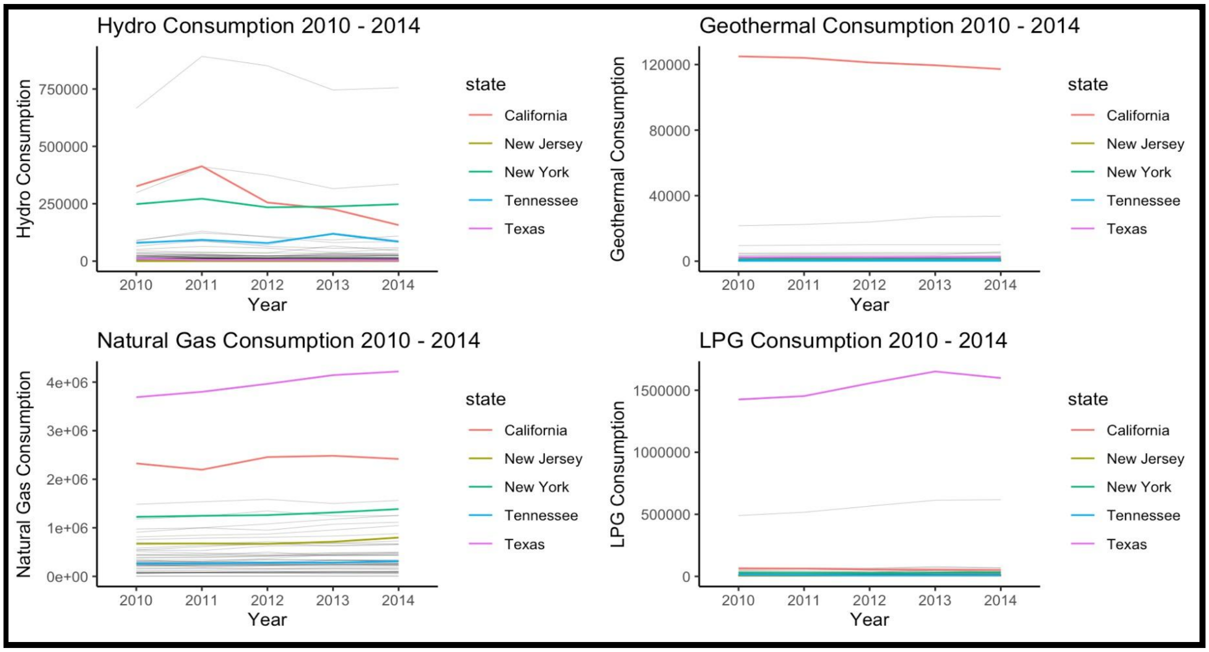


In the above graphs, we see the percent change in total energy expenditure year over year from 2010 to 2014. Energy expenditure represents the total amount of energy used by each state. The most noticeable state from the maps above is, once again, North Dakota and how it ranks near the top across the five years in the dataset. While it ranks 48th in population, it has had some of the highest increases in energy expenditure, production, and consumption. The mean percentage change of the total energy expenditure for each year is represented by a black dashed line. The biggest increase in mean production happened in 2011, while we saw only a slight increase in 2013 and 2014 with a decrease in 2012.

Yearly Consumption

From our dataset, we had variables that represented the totals for each energy producer for the years 2010 through 2014, so we needed to transform the data to fit the spaghetti plots below. We used the pivot_longer function in Rstudio to create a "Year" and individual consumption column to represent the consumption values for each given year. In the spaghetti graphs below, we decided to have five states represented (California, New Jersey, New York, Tennessee, and Texas.) This allowed us to focus on these five states and see how different sized states (landmass and population) will differ from each other. The most noticeable graph below is the geothermal graph and how far beyond California is compared to the other 49 states. It was touched on in the energy analysis above how California's massive geothermal consumption number was enough to add correlation to the GDP of the United States. The biggest surprise we had when running the analysis below was that South Dakota was the leader in hydro consumption in the years represented in the data. When we did more research, we learned that in the year 2020, renewable energy accounted for 83% of South Dakota's energy production and 50% of that was hydroelectric power. We can conclude that South Dakota was well on its way to finding alternative, renewable forms of energy in the years 2010 to 2014. In many cases we see California and Texas at the top of the United States, which was not surprising due to the sizes and population of each state.





Conclusion:

From our GDP analysis, there has been a huge jump in a city's GDP. We realize that California has the greatest GDP increase. Additionally, California also has the most GDP in the given year. Moreover understanding the trend in GDP increase or mean GDP made us question and want to analyze what factors really affect the GDP.

In this paper, we have analyzed the effect of different parameters on the GDP. For the population part, we can conclude three things. Firstly, population has a very strong correlation with GDP. Secondly, population increase, especially significant population increase is likely to drive the GDP growth and the amount of population increase positively influences the amount of GDP increase. Thirdly, the death rate and the net international migration rate can be regarded as two important factors to predict the GDP.

For consumption analysis, while we find four groups, including consumption, production, expenditure will impact GDP, the yearly changing analysis may be more complicated than the total amount demonstrated. However, we could still find some good indicators here. Generally, we believe more expenditure and more investment will increase GDP. Energy consumption is another large part here. We realized that more consumption will indicate a higher GDP. While some energy consumption will be little different, the general tendy is consistent here. While individual energy production shows a positive relationship with GDP increase, outliers play the most roles here. Thus, we will exclude it as one major impact here. Additionally, we realize that the price is not a good representation here.

Limitation:

1. Limited period. We only have five years of data. The impact of many factors on GDP takes longer to be reflected, such as the birth rate.
2. In this project, we analyzed the impact of various factors on GDP separately. This method of analysis will prevent us from knowing the impact of multiple factors on GDP.
3. Mystery of Death rate: From the correlationship, there is a significant negative relationship between death rate and GDP in 2011 and 2012. However, such correlation was not significant in

2013 and 2014. We don’t know if this phenomenon is temporary or permanent. In other words, we are not sure whether the death rate can be still regarded as an important factor to predict GDP.

4. Limited information about the extreme outliers. For example, when we analyzed the correlation between GDP and energy production. We found that geothermal energy production has the strongest correlation with GDP. However, in fact, only a few states produce geothermal energy. We need more information about geothermal energy to determine whether it can be regarded as a good indicator of GDP.

How to improve:

Based on the limitations, some ways to improve our research is by expanding our dataset to more years. This will help us validate all our findings and see if the same trends hold true for the rest of the years. The Mystery of Death and limited information about the extreme outliers can be solved through searching relative data. Moreover, a predictive analysis model can be created, provided more data is available, to help better understand the current GDP. We can use PCA to combine various variables and generate new variables to solve the problem of separate analysis.

Reference:
https://www.kaggle.com/lislejoem/us_energy_census_gdp_10-14/tasks