Implementing Bayesian Hypothesis To Develop Choropleth And Surprise Maps of GDP Per Capita In Relation To Energy Usage

CSC 47400 - Assignment 2

Nayma Labonna
Liana Hasan
Haoyuan Wu (Howard)
Deepankar Chakraborty

Objective

The objective of this assignment is to implement the Surprise Map model developed by Micheal Correll and Jeffrey Heer and use a dataset to come up with our own surprise map. In order to create the surprise map we used the Bayesian model to come up with a hypothesis and then used arithmetic calculations to generate the signed surprise values. By using Python for visualization, we created the choropleth maps to visualize the initial data, our hypothesis, and the final surprise map.

Dataset

Our dataset consists of the GDP per capita of each individual nation in the world in 2019. GDP or Gross domestic product is the total monetary value of all finished goods and services produced within a country's border[1]. Economic growth or the GDP of a country has been historically strongly correlated with the total energy consumption of a nation, because, as a country grows economically, it requires more energy to generate goods and services. In our surprise model, we wanted to see if higher GDP correlates with higher energy usage. Hence, our evidence is the per capita energy consumption across countries around the world in 2019 [2]. Based on the evidence data, we want to test our hypothesis using a surprise map and discover any outliers or surprise values.

Bayesian Model

We want to implement Bayes theorem in order to come up with a hypothesis that we can test in our visualization. Bayes theorem is used to find the probability of a hypothesis with given evidence.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

At first we constructed a model space, ***A,*** comprising our dataset of GDP per capita. Then we generated an initial set of prior beliefs or hypotheses, which is "***the GDP of a country is strongly correlated with its energy consumption. So the higher a country's GDP per capita is, the higher its energy consumption.***" Finally, we collected data, ***B***, which is the energy consumption per capita in 2019 of each individual nation to come up with the Bayesian model to test our hypothesis. We can write the Bayesian surprise model for our experiment as follows:

$$p(gdp\ p.c.\ |\ energy\ consum.\ p.c.) = \frac{p(energy\ p.c.|gdp\ p.c) \times p(gdp\ p.c.)}{p(erergy\ consum.|\ p.c.)}$$

[Note: ***p.c.*** = per capita***,*** *consum****.*** = consumption]

Setting Up Data

First, we uploaded the two datasets we planned on using on to a Github repository. The first dataset is labeled 'GDP per Capita' and the second dataset is labeled 'Energy per Capita'. These datasets are shown visually on a map of the world using dark and bright colors to indicate how high or low the GDP and Energy is in different countries all over the world. Our Surprise map will have a similar format.

Using the raw Github links, we were able to access the datasets as CSV files using a Python software library called 'pandas' which is a software library used for data manipulation and analysis. Using this library, we used the `read_csv()` function to access and read the datasets from Github.

Then, we used Python to clean up the datasets. We didn't have access to the data of energy per capita of some countries, so for the 'GDP per Capita' dataset we used the `split()` function to clean the countries and drop certain columns we didn't need. We also noticed that the 'Energy per Capita' dataset had multiple years even though the GDP dataset contained data only from 2019. So we filtered the Energy dataset using the `query()` function to be only for the year 2019. Then we renamed the label 'Entity' in the Energy dataset to be labeled as 'Country' using the `rename()` function, since the Entity column contained the countries. Then we dropped the columns 'Year' and 'Code' using the `drop()` function since they weren't necessary for our objective.

Then, for the Surprise Map we combined both of the cleaned datasets which we labeled 'energy' and 'gdp' using the `merge()` function in this line "`combined_df = pd.merge(gdp, energy, on='Country')`". This combined dataset contains the energy and GDP per capita.

Method

A surprise map shows how much a value is off from the expected value: too high means positive, too low means negative. The surprise value can be calculated as:

$$\text{Surprise} = \frac{Expected - Actual}{Expected}$$

Since our hypothesis is as a country uses more energy it has a higher GDP, we calculated the expected value by finding the average of $\frac{GDP}{unit\ energy\ use}$.

$$P\left(\frac{GDP}{unit\ energy\ use}\right) = \frac{1}{n}\sum_{i=1}^{n} P_i\left(\frac{GDP}{unit\ energy\ use}\right)$$

This allowed us to create the formula for out surprise value $= \dfrac{\frac{1}{n}\sum_{i=1}^{n} P_i\left(\frac{GDP}{unit\ energy\ use}\right) - P(GDP)}{\frac{1}{n}\sum_{i=1}^{n} P_i\left(\frac{GDP}{unit\ energy\ use}\right)}$.

The surprise value for each country should be slightly above or below 0, because 0 means the actual GDP is the same as expected which is not surprising at all. But a too large or too small value can be considered as an outlier. We can determine if our hypothesis is true or false by observing the trend/color of the surprise map.

Observations and Analysis

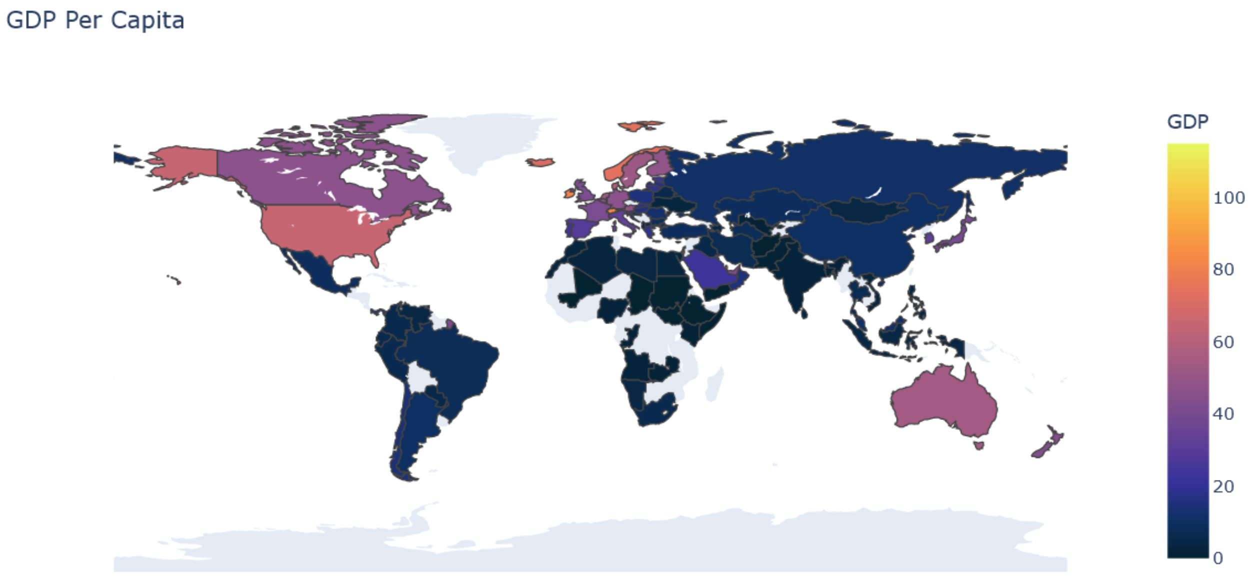

Figure 1: This map shows the GDP per capita of different countries.

According to this map, we can see that most countries have low GDP per capita, which are represented in dark blue, and a few countries such as the USA and Australia have relatively high GDP per capita since they are a reddish color, and countries like Ireland which are orange have the highest GDP per capita. Countries that do not fall within the color scale, like the few countries in Africa and South America, are countries that we do not have data for.

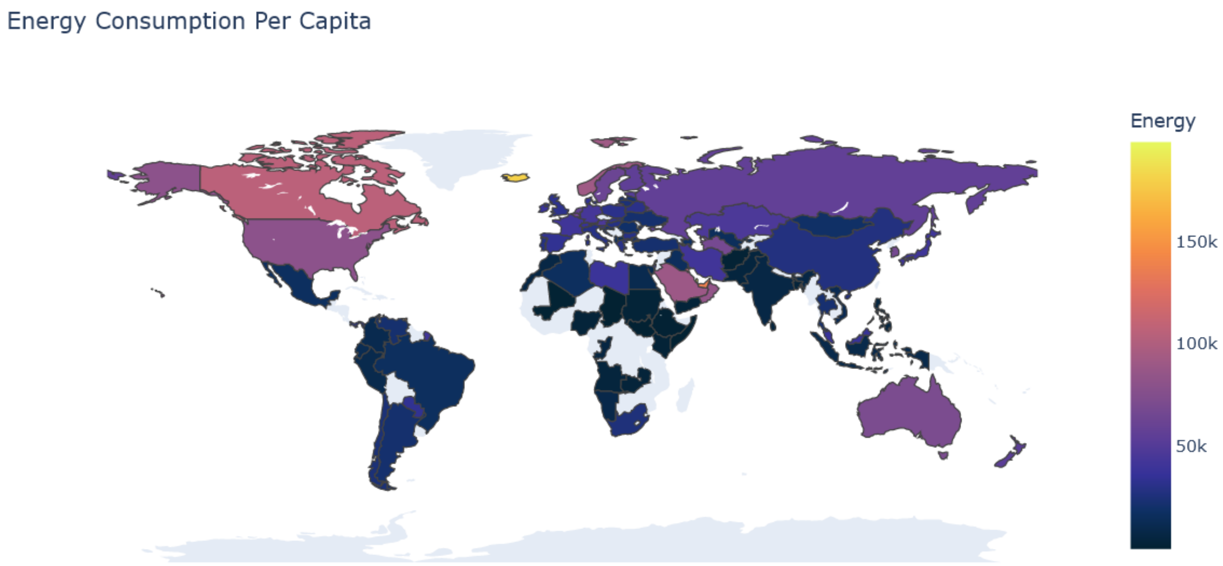

Figure 2: This map shows the energy consumption per capita of different countries.

In figure 2, we can also see that most countries do not use a lot of energy as seen by the countries in dark blue. Some countries, like the USA, Canada, and Australia, have a relatively high consumption of energy which is shown as light purple, and countries like Iceland have a very high energy consumption rate per capita and can be seen in the map as yellow.

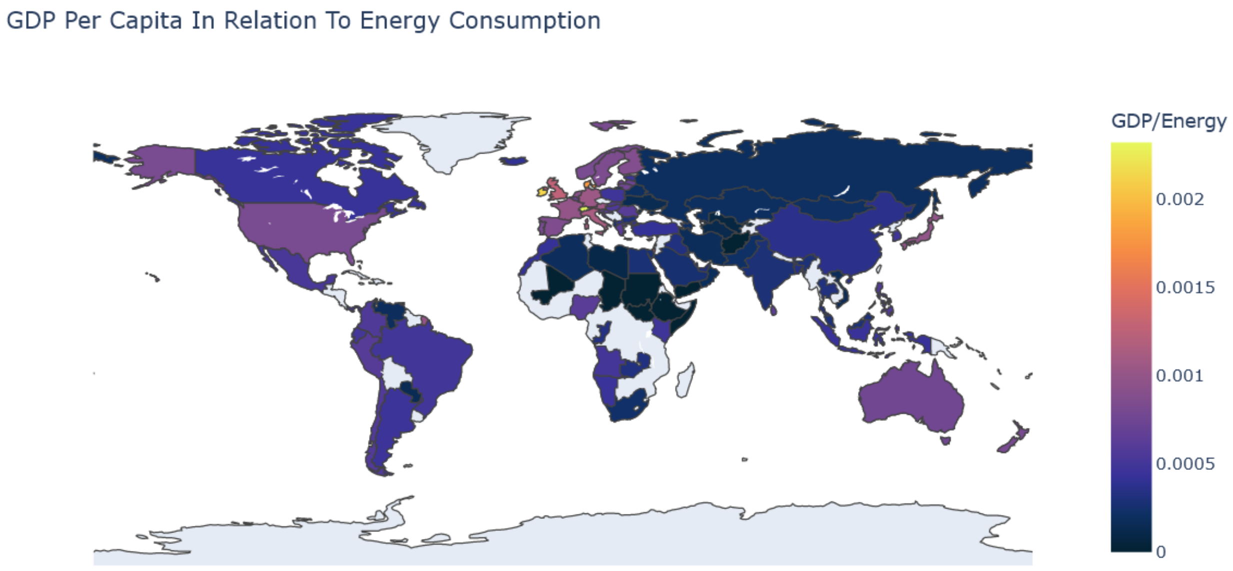

Figure 3: This map shows the GDP per Capita in relation to energy consumption per capita of different countries.

This map allows us to see which countries have a high or low ratio of GDP to energy consumption per capita. Countries that are light purple like the USA and Australia have similar ratios and both have an average value of GDP and energy usage. This will be used to calculate our expected value, which is essentially the average ratio for all countries.

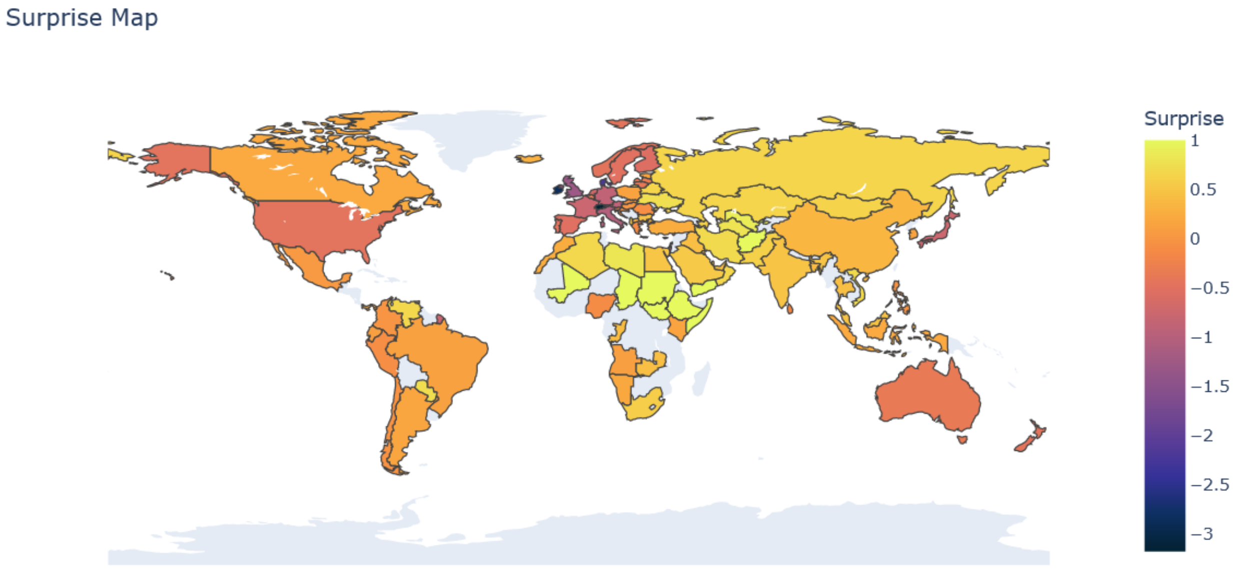

Figure 4: The surprise map.

As we can see here in the surprise map, most of the countries have the color around 0 which is the orange color. This means that a majority of our data falls within our expected value for GDP in relation to energy consumption per capita, so it is significant that our hypothesis is true. Countries that are dark blue or black have high GDP but low energy consumption. These countries do not align with our hypothesis and are considered outliers. This can be seen with Switzerland, which has a high GDP, but surprisingly low energy usage, which is why it has a negative surprise value and is shown as black. The bright yellow outliers have both low GDP and low energy usage, which can be seen in some of the African countries such as Sudan and Chad.

Link to the code on Github Repository: https://github.com/lianahasan/CSC-474-Assignment-2

Team Member Roles

We met frequently and did the work uniformly. Here is a rough outline of how we divided our tasks and came up with the final outcome:

Deepankar: Found datasets, implemented bayesian theory to come up with the hypothesis, and created choropleth map for gdp per capita

Liana: Created github repository and google collab notebook and created choropleth map for energy consumption per capita

Howard: Came up with calculation and formula for surprise map and created choropleth map for gdp in relation to energy consumption per capita

Nayma: Cleaned both datasets and created the final surprise map using Python.

References

1. Gross Domestic Product (GDP)Definition, https://www.investopedia.com/terms/g/gdp.asp
2. Hannah Ritchie and Max Roser (2020) - "Energy". Published online at OurWorldInData.org. Retrieved from: https://ourworldindata.org/energy-production-consumption