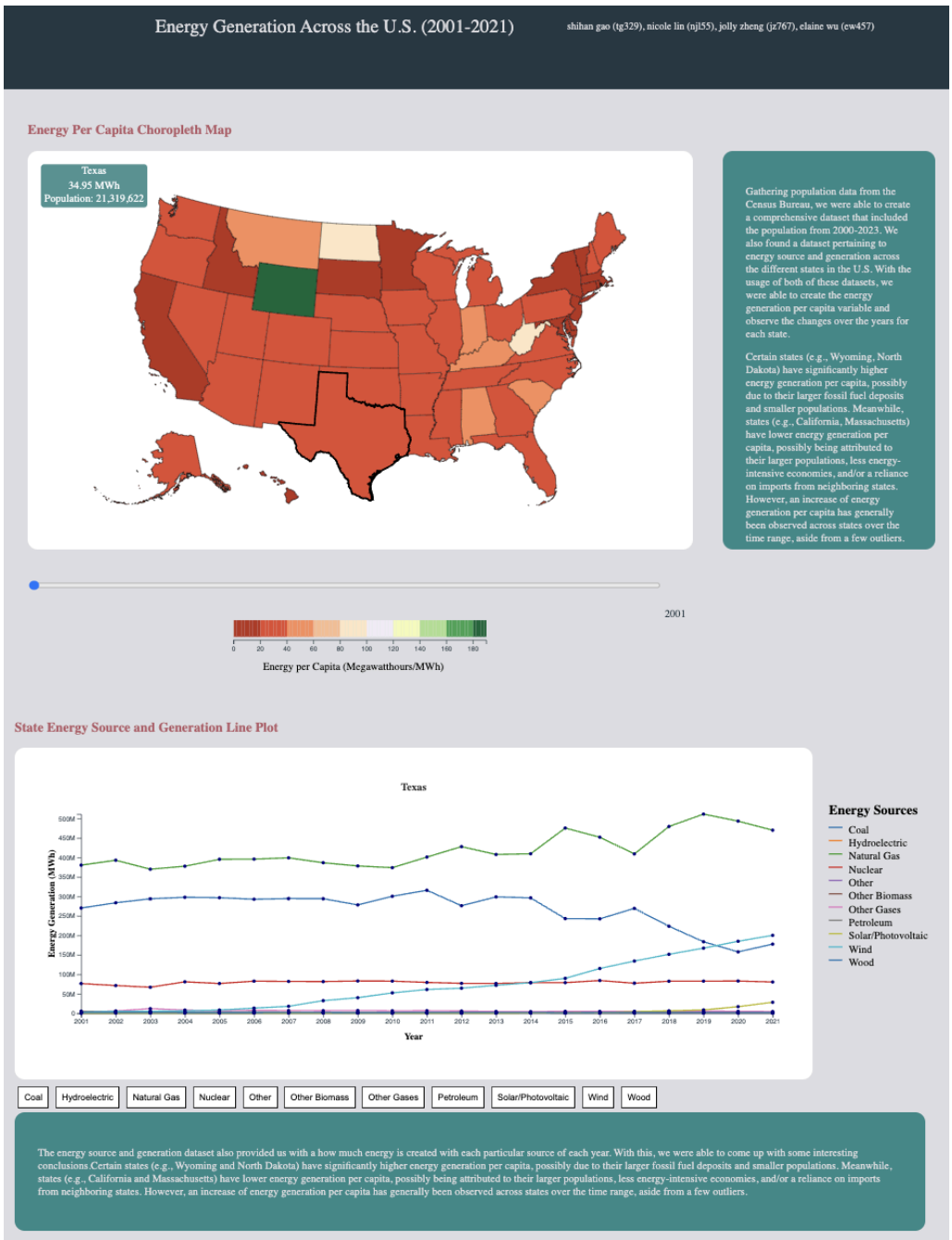
Shihan Gao (tg329), Nicole Lin (njl55), Elaine Wu (ew457), Jolly Zheng (jz767)
November 13, 2024
INFO 3300

INFO3300 Project 2: Technical Write-Up

NOTE:
VIEW GRAPH VISUALIZATION HTML IN FULL-SCREEN FOR BEST EXPERIENCE



Overview of the Dataset:

Our project used a total of 5 datasets: three about population, one about energy generation and sources, and two about the U.S. states.

- Population Datasets:
  The three population datasets were taken off the United States Census Bureau website and were originally in a xlsx (Excel) format. The general format of the file itself can be described in the columns offered:
    - Column Location: This was spread across the geographic area which included the 50 states, the District of Columbia (Washington D.C.), and Puerto Rico as well as overarching locations like the United States, Northeast, Midwest, South, and West.
    - Years: Some years have firm numbers which is a separate column, typically at the start of the year columns group. Then some years are just estimates. At the end, there is a form number of population estimate, at the end of year columns group.

  Resource Links:
    - Data to get 2000-2009: Annual Population Estimates
    - Data to get 2010-2019: Tables; Annual Estimates of the Resident Population for the Nation and States
    - Data to get 2020-2023: First Excel under category, Population Estimates, Population Change, and Components of Change
- Energy Source & Generation Dataset:
  The energy source & generation dataset was taken from Kaggle. The dataset had a total of six variables which are the following:
    - YEAR: year the data was collected from
    - MONTH: month the data was collected from
    - STATE: what state the data pertains to
        - written in State Code (two letter code of a state)
    - TYPE OF PRODUCER: type of producer of this certain energy
        - there is a total of 6 types which are the following:

Combined Heat and Power, Electric Power	Combined Heat and Power, Industrial Power	Electric Generators, Electric Utilities
Electric Generators, Independent Power Producers	Electric Generators, Independent Power Producers	Combined Heat and Power, Commercial Power

    - ENERGY SOURCE: type of sources
        - there is a total of 13 types of sources of energy:

Coal	Hydroelectric Conventional	Natural Gas	Nuclear
Petroleum	Wood and Wood Derived Fuels	Solar Thermal and Photovoltaic	Wind

Pumped Storage	Geothermal	Other Biomass	Other Gases
Other			

    - GENERATION (Megawatt-hours): generation of energy in megawatts-hour within the given constraints of state, month, date, source and generation type
- U.S. States
    - States Name file
      This file was provided with the energy source & generation dataset on Kaggle. The columns are listed as follows:
        - State: State's full name
        - Abbrev: abbreviation of the state
        - Code: Two Letter Code for States
    - U.S. Topojson file
      This file was found through NPMJS. We used the file that had the states information retained to the topojson file. The state information are further declared as follows:
        - id: two-digit state code
        - properties.name: state full name
        - properties.code: state two-letter code
        - properties.abbrev: state abbreviation
        - properties.type: legal type of entity, etc
            - can be the following: "state"
            - "federal district"
            - "insular areas

Filtering and Preprocessing the Data

There were a couple of things that were done to get the files ready to be used. It can be described in the following subsections:

- Populations
  For each population dataset, it was originally in the form of an Excel file. However, we wanted CSV files to work with. As a result, we converted the Excel file into a CSV format which caused a couple of issues with how the files translated over. Therefore, our steps of preprocessing were as follows:
    1. Convert Excel to CSV format
    2. Drop all rows before the first state (Alabama)
    3. Drop the line between Wyoming and Puerto Rico
    4. Drop all rows after the Puerto Rico

  With all the rows required in the csv files, we started renaming the columns. This was done differently across each file but followed the same logic.
    1. Rename the first row to state
    2. For start year columns we rewrite it as a form of "Estimate #".
        a. Years 00-09: Second column is "Estimate1"
        b. Years 10-19: Second column "Estimates1". Third column is "Estimates2"
        c. Years 20-23: Second column is "Estimate1".

3. In the next few columns, we would see the actual estimate of population so we label the columns as the years accordingly.
    a. Years 00-09: Next few column names are as follows: "2000", "2001", "2002", "2003", "2004","2005", "2006", "2007", "2008", "2009”
    b. Years 10-19: Next few column names are as follows: "2010", "2011", "2012", "2013", "2014", "2015","2016", "2017", "2018", "2019"
    c. Years 20-23: Next few column names are as follows: "2020", "2021", "2022", "2023"
4. If there is estimates following the years, we label these as estimates as well:
    a. Years 00-09: Next few columns are named as follows: "Estimate2", "Estimates3"
    b. Years 10-19: Next few columns are named as follows: "Estimate3", "Estimates4"

In the next step, we drop all columns that are named with the Estimate # as it only provides extra information. We only want to keep the estimates for each year. Lastly, when we converted to CSV, all the States had a ‘.’ in front of them. We stripped the ‘.’.

We then exported the data frame to be a csv file, which is in the following folder locations:

- Years 00-09: filters/populations/00-09/00-09.csv
- Years 10-19: filters/populations/10-19/10-19.csv
- Years 20-23: filters/populations/20-23/20-23.csv

However, all these years are currently in separate data frames at the moment. We wanted to put them in one data frame for convenience in making the data visualizations. As a result, we merged all three of these data frames and exported this to a CSV file which resulted in the following file: pop_all_years.csv.

- Energy Source & Generation Dataset

  There was also data preprocessing and filtering for the energy source and generation dataset. We created two CSV files from this dataset which are source_years.csv and producer_years.csv. For both dataset, we start by reading the CSV file both states and the general energy source and generation dataset.

  producer_years.csv

  1. From energy dataset, filter out rows where the energy source is 'Total' to focus on specific energy types.
  2. Aggregate energy data by state, year, type of producer, energy source, and month. Summing up the energy generation values for each group.
  3. Further aggregate the data by removing the 'MONTH' level, summing up energy generation across all months for each state, year, producer type, and energy source.

  source_years.csv

  1. Read CSV file for both states and the general energy source and generation dataset.
  2. Filter out rows where the energy source is 'Total'. This ensures the focus is on individual energy sources rather than overall totals.
  3. Group the filtered energy data by state, year, and energy source, and calculate the total energy generation for each group.
  4. For the sake of simplicity in name, we renamed the following cells that have the following strings:
      a. 'Wood and Wood Derived Fuels' => 'Wood'
      b. 'Hydroelectric Conventional' => 'Hydroelectric'

c. 'Solar Thermal and Photovoltaic' => 'Solar/Photovoltaic'

Now that we have the data that we would like, the next step is uniform for both. We start by renaming the 'Code' column in the state DataFrame to 'STATE' for consistency in merging. Then, merge the state data with the aggregated energy data on the 'STATE' column. Next, we perform a left join to retain all states even if there is no matching energy data. Drop unnecessary columns 'Abbrev' (state abbreviations) and 'STATE' (state codes) from the merged data as we only want the full name of the state. Lastly, for each of the new datasets, save the final processed data to a new CSV file.

Overview of Design Rational

CSS AND HTML**:**

The graphs are positioned in a way so that it is easy for the user to understand at first glance. The main thing that they should focus on is the map, and the legend and slider are in an easy to spot location. The eyes are then drawn to the line graph once a state has been selected, and then down to the buttons, which follows the natural flow from the top to bottom of the page.

CHOROPLETH MAP DESIGN RATIONALE:

The Choropleth Map uses a color scale with red representing low energy per capita and green indicating higher energy per capita. Red often signifies something negative or low, while green is typically associated with positive or high values. A horizontal year slider placed at the bottom of the map which allows for easy access to update the map. The year displayed to the right of the slider provides feedback about which year's data is being viewed. A legend is included for easy reference of the color scale, and a tooltip appears when hovering over a state, providing contextual information such as the state name, energy per capita in MWh, and total population. The tooltip fades out smoothly (during mouseout) and is positioned at the top left to avoid overlapping with any map states. When a state is clicked, a heavier, black outline (the "momesh") is drawn to indicate that it has been selected.

LINE PLOT DESIGN RATIONALE:

For each of the 50 states, when clicked, it displays a line graph of the states' energy generation for each of their sources over the years. If no state has been clicked, it displays the default message in the center, provides the user with feedback on what actions to take next, improving the user experience and preventing confusion. This infographic would be helpful since it allows the viewer to get into the nitty gritty of the energy generation of each state. In this case, they would be able to see the top sources that are generating energy. The graph is purposely made to be a bit wider for aesthetic fitting purposes.

In order to indicate which state is selected, the state name appears at the top of the graph. The axes are drawn for both the X and Y axis. The Y axis specifies what the measurement is (with the full measurement name in the description, since it would look a bit crowded to include the full name). The numbers on the Y axis are dynamic and change per state since some of the states energy generation differs by a few million, and therefore some of the line graphs would not be as legible. Furthermore, the units are given as "88m = 88 million," shortened for legibility. The circles represent individual data points, and change color when hovered, providing immediate feedback to the user. The text label displays the energy generation value for that particular source given that specific year. Furthermore, the labels next to the data points dynamically adjust their position to prevent overlap with the graph's borders and ensure legibility. This improves user experience, especially when data points are near the edges of the graph.

Interactivity: click and view from choropleth (50 line graphs for 50 states), mouse hover, data point views, buttons

Overview of Interactive Elements & Their Design Rationale

CHOROPLETH MAP INTERACTIVITY RATIONALE:

The Choropleth Map is interactive and linked to the line plot, allowing users to see how energy production per capita varies across states over time. The map includes a year slider, enabling users to select a year between 2001 and 2021, with the map updating dynamically to reflect the chosen year's data. A tooltip appears when the user hovers over a state, displaying information such as the state name, energy per capita in MWh, and total population. The tooltip fades out smoothly and is fixed at the top left corner to avoid overlapping with the map. Additionally, when a state is clicked, a black outline (the "momesh") is drawn around the selected state, serving as a visual indicator and linking the map selection to the line plot.

LINE PLOT INTERACTIVITY RATIONALE:

There are small circles on the line graphs for clarity. When the user hovers their mouse over the small circles, it would turn a different color and showcase the source and the energy generation amount. This is so that there would be clarity on what exactly the source is instead of having the user find the line color in the legend and tracing the line to the Y-axis and approximating the amount. Furthermore, the circles are small, but noticeable. If they were too big, it would distract the viewer from the actual lines themselves, which are important for showing the trends of the energy generated for each of the sources.

Our line graph also includes functionality for toggling visibility of the different energy sources. Once a state is clicked on, buttons appear below the line graph for each of the corresponding energy sources available in the data set. Initially, all of the energy sources available are visible in the line graph, but when a button is clicked, the corresponding line is no longer visible on the graph and the legend. Thus, any combination of the energy sources can be viewed, whether it is several lines or only one, which is essential when attempting to visualize the data in certain states, as several lines tend to overlap near the x-axis for the majority of states. Take the line graph for Massachusetts as an example. The majority of the lines (aside from coal, natural gas, and nuclear) are difficult to see individually as they all crowd near the x-axis. The visibility toggle function allows for the user to directly compare certain energy sources that would otherwise be lost if every energy source was visible. This can be especially useful for comparing interesting sets of data, such as the line representing petroleum in the Massachusetts-specific line plot where it starts high and eventually dips towards the x-axis. This line is easily lost among all of the lines representing each of the energy sources once it dips towards the x-axis, but the toggle functionality allows the user to directly compare the energy source’s decline with other energy top or lesser-used sources. Therefore, our interactions toggling line visibility allows the user to focus on specific energy sources and remove any irrelevant information.

The Story

Our visualizations reveal the balances and imbalances between states’ energy generation and the demands of their population while highlighting shifts in energy sources over time. Our choropleth map visualizing energy per capita shows the variation in energy generation per person across different regions over a time period of 2001 to 2021. States are shaded in different colors, with lighter hues indicating higher energy generation per capita. Our map shows certain states, such as Wyoming and North Dakota,

stand out with significantly higher energy generation per capita, which could potentially be explained by their energy-intensive industries and smaller populations. Meanwhile, states like California and Massachusetts have much lower energy generation per capita, possibly being attributed to their larger populations, greater energy efficiency initiatives, or less energy-intensive economies. These differences in energy generation per capita can also be possibly explained by local energy resources. States with larger fossil fuel deposits (e.g., West Virginia is a leading producer of coal in the U.S.) tend to generate higher amounts of energy relative to their population size, while more densely populated states may produce less energy, resulting in a reliance on imports from neighboring states (take Massachusetts as an example, the third most densely populated state in the U.S. that is also a top power-importing state). Interestingly, using the slider attached to our choropleth map demonstrates an increase in energy per capita over the years as more states become a lighter hue, particularly in states with energy-intensive industries. This may be due to the expansion of energy infrastructure, greater extraction of energy resources, or a push for energy independence in certain states. Notably, a few states, such as West Virginia and Alaska, seem to have declined in terms of energy consumption per capita, which is noteworthy in the context of the nationwide shift towards sustainability and renewable energy sources and can be further explained with our second visualization.

Our second visualization presents energy generation partitioned into different types of sources, which provides further insights into how energy generation sources have shifted over the past two decades. A nationwide shift to renewable energy sources is demonstrated as states' usage of coal, a previously popular fossil fuel, have generally declined while there is a slight uptick in sustainable sources of energy, like wind and hydroelectric. There is a stark contrast in energy generation per capita and the sources of energy across the past few years in states traditionally known for coal-mining (and as a result, mostly rely on coal as a source of energy). If we zoom in on these specific states, primarily Wyoming and West Virginia, there is a significant change in color over the years. Looking at their state-specific line graphs showcasing different energy sources, there is a substantial decline in their primary source of energy, coal, while there is only a slight increase in renewable energy sources, like wind. Although there is a noticeable decrease in coal as an energy source, other fossil fuels, particularly natural gas, tend to still be widely used and in some states, increasing even. Several states, like Connecticut, Virginia, Alabama, and Arizona, demonstrate a broader national trend where natural gas has gained prominence as an energy source, most likely due to its lower carbon footprint compared to coal and its economic advantages. Natural gas seems to have become a key transitional energy source, playing a larger role in states that have phased out or significantly reduced coal use, like Florida. The increased adoption of natural gas in many states is indicative of shifting energy policies that aim to balance energy needs while still considering the environment.

Combining these visualizations together gives a broader understanding of how energy generation patterns are shifting both geographically and technologically. The choropleth map shows which states are producing the most energy per capita, often corresponding to states with energy-intensive industries or vast energy resources; meanwhile, the line plot delves deeper into the sources of that energy, showing that states are increasingly moving toward natural gas and renewable sources as they phase out coal. However, the rise of natural gas suggests that while states are making progress in reducing reliance on coal, the shift to fully renewable energy sources is still ongoing. Natural gas, although cleaner than coal, still is a fossil fuel that contributes to greenhouse gas emissions, meaning that these states are not yet fully embracing the zero-carbon energy future that many are striving for. The increase in energy generation per capita across many states, as suggested by the map, raises questions about the future of energy policy. As states,

like Louisiana and North Dakota, continue to ramp up energy production, primarily due to industrial energy demand, they face the challenge of balancing economic growth with sustainability. This dynamic reflects the broader challenges facing the energy sector today—meeting growing energy demand while transitioning to more sustainable, environmentally friendly energy systems. Our data visualizations tell a story on how the United States' energy landscape is evolving and what steps may be needed to ensure a sustainable, reliable energy future.

Outline of Team Contribution

The form we had previously filled out outlined the following roles: visualization programmer, visualization designer, project manager, and report writer. While we all did not formally take on a singular role specifically, we took on each role when necessary. We can say our visualization was developed in the following stages:

- Choosing Final Dataset
- Designing Data Visualization
- Implementing Data Visualization
- Writing Technical Report

We wanted to create a cohesive and strong story from our data visualizations which led us to constantly go back to the drawing board on what dataset we wanted to use and how to use it. It took us around **five** hours to finally decide on using the energy generation and sources dataset. It then took us another **hour** to find the other population dataset that would make the filtering and preprocessing the most standard throughout. As for the filtering and preprocessing itself, it took us around two hours to figure out how to filter the data in a way that best suited the story we wanted to present.
As for designing the data visualizations and finding overlapping interactivity between them, it took us around **four** hours to deliberate between several different designs that convey different components to the story we are trying to show. We went back to the drawing board and worked through creating a coherent story throughout our visualizations. We went through many iterations of considering what other variables that would best demonstrate our story. We collaborated and critiqued each design and decided on a couple that would best convey our story.
Coding the visualization took the most time, which was approximately **twenty–one** hours, as each of us worked on separate parts of the visualization and we went through several iterations of our visualizations before settling on a final appearance.

In the end, it collectively took us about **seven** hours to write up the technical report, to ensure our design rationale and story were properly conveyed in text.

Shihan Gao: I focused on designing and implementing the Choropleth Map, along with its interactivity. I also developed the year slider functionality, which updates the map's data in real-time. Additionally, I implemented a legend that provides a clear reference for the color scale of the map. Finally, I worked with Nicole on the CSS to improve the overall layout of the project, ensuring that it is aesthetically cohesive.

Nicole Lin: As a team we collaborated on generating ideas and looking for viable and good datasets to use. After finding the energy generation dataset, we all talked through how to use it and what visualizations we could create. I created the Line Graph, and implemented its interactivity. I also created

the dynamic legend for it. I further implemented the functionality of clicking on each of the states on the map to produce a new scatter plot with that particular state's data. Furthermore, I worked on the HTML layout and designed with CSS to make the project look cohesive and aesthetically pleasing. Afterwards, I wrote the HTML/CSS Layout Design Rationale, Design Rationale for the Line Graph and the Interactivity Rationale for the Line Graph sections of the technical report.

Jolly Zheng: We worked together on brainstorming ideas and finding datasets. In particular, after finding the energy generation dataset, I worked on how to use it in conjunction with that dataset. As a result, I found the three population datasets. In addition, I worked on cleaning all the datasets and adding comments to the code inside both the python and index.html files. I also wrote the Overview of the Dataset, Filtering and Preprocessing of the Data, and Outline of Team Contribution sections of the technical report.

Elaine Wu: As a group, we brainstormed visualization design ideas, what story we wanted our visualization to tell, and searched for viable data sets together. I helped implement the basic skeleton of the choropleth map, and I implemented the toggle buttons and the corresponding functionality for the line plot. For the technical report, I wrote the Line Plot Interactivity Rationale section and analyzed the data and visualizations in order to write The Story section.