Eric Mach

DX699 Mod B: AI For Leaders

Semester 2

Milestone 3

Section 1: Description of Project

For Milestone 3, the three datasets that were selected relate to renewable energy trends across the U.S. and the world over time. All the selected datasets span a period of at least the beginning of the 21st century until 2022, with one dataset spanning back to 1965. All datasets track energy generation, breaking down the types and methodology that each country/state utilizes as well as how those trends may have changed over time. Collectively, these datasets have the potential to provide significant insight into the global shift toward renewable energy and policy development to meet sustainability goals. Countries with specific scalable successes in renewable energy adoption could be used as beacons to guide developing nations in leveraging their natural resource advantages. From a policy perspective, surging energy generation may highlight the need for installed capacity expansion to accommodate the increased demand, providing incentive for future investment into both existing energy architectures and emerging alternative energy technologies. In the exploratory data analysis below, the datasets will be examined for both their suitability for further analysis and any potential insights that could be gained from initial visualizations. As the datasets are treated as a collective set, each step of the exploratory data analysis will detail the combined conclusions of all datasets, with specific examples shown from each one as necessary.

Section 2: Exploratory Data Analysis (Preprocessing + Univariate)

In the first step of exploratory data analysis, each dataset was preprocessed, checking for missing values and initial statistics of the distribution of the data. Overall, none of the datasets exhibited concerns that would deem them unsuitable for analysis at this point. In fact, getting a broad look at the breakdown of the dataset format helped in shaping a plan as to how to further analyze the dataset. For example, one of the global datasets (Har-var, 2024) comes from the International Renewable Energy Agency (IRENA), offering comprehensive energy production data on various regions, sub-regions, and countries from 2000 to 2022. The numerical metrics include electricity generation and installed electricity capacity for differing energy technologies. An important distinction to note is that electricity generated is a measure of energy produced over a period of time while installed electricity capacity is a measure of power which describes the maximum potential power output at any particular moment of a generating facility under optimal conditions.

```
import chardet
file_path = 'IRENA_RenewableEnergy_Statistics_2000-2022.csv'

with open(file_path, 'rb') as f:
    result = chardet.detect(f.read())

irena = pd.read_csv(file_path, encoding=result['encoding'])

irena.head()
```

	Region	Sub-region	Country	ISO3 code	M49 code	RE or Non-RE	Group Technology	Technology	Producer Type	Year	Electricity Generation (GWh)	Electricity Installed Capacity (MW)
0	Africa	Northern Africa	Algeria	DZA	12	Total Non-Renewable	Fossil fuels	Natural gas	On-grid electricity	2000	24585.0	5459.01
1	Africa	Northern Africa	Algeria	DZA	12	Total Non-Renewable	Fossil fuels	Natural gas	On-grid electricity	2001	25781.0	5455.50
2	Africa	Northern Africa	Algeria	DZA	12	Total Non-Renewable	Fossil fuels	Natural gas	On-grid electricity	2002	26994.0	5891.01
3	Africa	Northern Africa	Algeria	DZA	12	Total Non-Renewable	Fossil fuels	Natural gas	On-grid electricity	2003	28619.4	6013.24
4	Africa	Northern Africa	Algeria	DZA	12	Total Non-Renewable	Fossil fuels	Natural gas	On-grid electricity	2004	30312.0	6305.24

Figure: IRENA Dataset Preview

As seen in the code above, the IRENA examines energy data on a global scale with a level of granularity ranging from continental to country-specific. Electricity generation is measured in gigawatt-hours and electricity installed capacity is measured in megawatts. Aside from region demographics, the data is also split by producer type and technology types. This gives further insight into how each country has leveraged different types of renewable energy and non-renewable energy over time. The describe function below details the distribution of the numerical data and the number of unique technologies utilized globally, including the most common technology, which is unsurprisingly fossil fuels/oil.

```
irena.describe()
```

	M49 code	Year	Electricity Generation (GWh)	Electricity Installed Capacity (MW)
count	35193.000000	35193.000000	3.519300e+04	3.159400e+04
mean	420.592618	2012.283437	1.440929e+04	3.452455e+03
std	252.192524	6.499073	1.148906e+05	2.715673e+04
min	4.000000	2000.000000	-5.874050e+02	1.000000e-03
25%	203.000000	2007.000000	1.168300e+01	5.000000e+00
50%	404.000000	2013.000000	2.002570e+02	5.890000e+01
75%	634.000000	2018.000000	2.383631e+03	6.325750e+02
max	894.000000	2022.000000	5.220700e+06	1.155325e+06

```
irena.describe(include='object')
```

	Region	Sub-region	Country	ISO3 code	RE or Non-RE	Group Technology	Technology	Producer Type
count	35193	35193	35193	35193	35193	35193	35193	35193
unique	5	17	224	224	2	10	19	2
top	Europe	Latin America and the Caribbean	United States of America (the)	USA	Total Renewable	Fossil fuels	Oil	On-grid electricity
freq	10632	7134	531	531	22185	10359	5658	28675

Figure: IRENA Dataset Descriptions

The counts remain consistent across all columns, except for some missing values in terms of electricity installed capacity, but not enough to derail an analysis. The other two datasets also exhibited very low, if any, missing values. However, they were all formatted similarly with this schema, with columns tracking years, regions, energy generation, and type of production to

varying degrees of detail. With all this data aggregated together without grouping by year, region, or technology, the general statistics are heavily skewed as seen in the electricity generated and installed capacity metrics. Medians are in the power of 10 or 10^2, while the maximums are to the power of 10^6. This is because energy generation data from 2000 from an underdeveloped country for example is being grouped with 2022 energy generation data from a developed country. This comparison issue with this key metric would be addressed in bivariate analysis.

When diving into univariate analysis, it was hard to gain much insight at all due to the aforementioned formatting of all three datasets. The skew of the electricity generation can be visualized below from the second global dataset (Hossainds, 2023). Similar to the IRENA dataset, this dataset breaks down total world renewable energy generation to country-level specifics, with separation between geo biomass energy production, solar generation, wind generation, and hydro generation in energy units of terawatt-hours. Note that it does not include fossil fuels or other non-renewable energy sources like the IRENA dataset does.

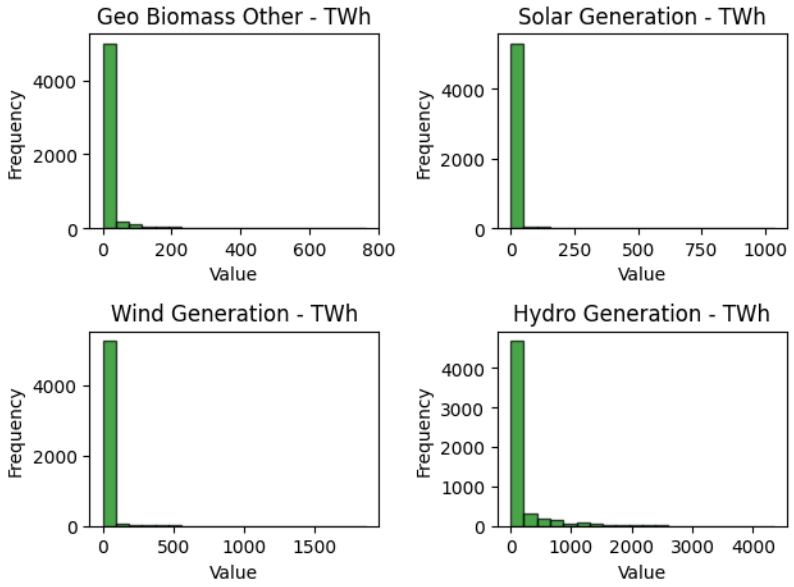

Figure: Renewable Energy Worldwide Dataset Histograms

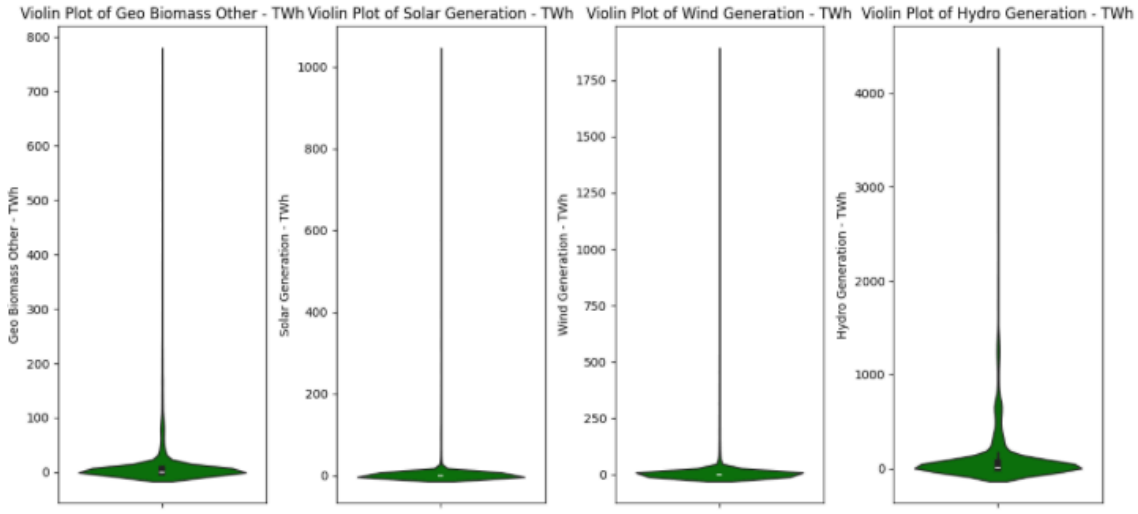

Figure: Renewable Energy Worldwide Dataset Violin Plots

As is evident in these histograms and violin plots, there is a heavy concentration towards the low end for all energy generation types. This trend was apparent across all datasets and was to be expected when looking at this metric from purely a univariate standpoint. Over recent decades especially, renewable energy and energy production has scaled dramatically; nearly at an exponential rate to meet increasing demand in this age of rapid technological innovation. As a result, recent years where the world has become much more efficient at harnessing renewable energy skews the dataset as a whole since only a few decades earlier, energy production was understandably much lower. While not necessarily insightful, seeing the scale of the skew in these univariate graphs was impactful in realizing the quickness in which the world has expanded its energy production capabilities.

Section 3: Exploratory Data Analysis (Bivariate)

Looking at these datasets from a bivariate perspective unlocks much more information regarding relationships of these variables as it pertains to time. Now, energy generation trends as it pertains to certain technologies, certain regions, and even certain countries can be examined. For

example, in the IRENA dataset, the world trends of electricity generation and electricity installed capacity can be visualized grouped by different technologies.

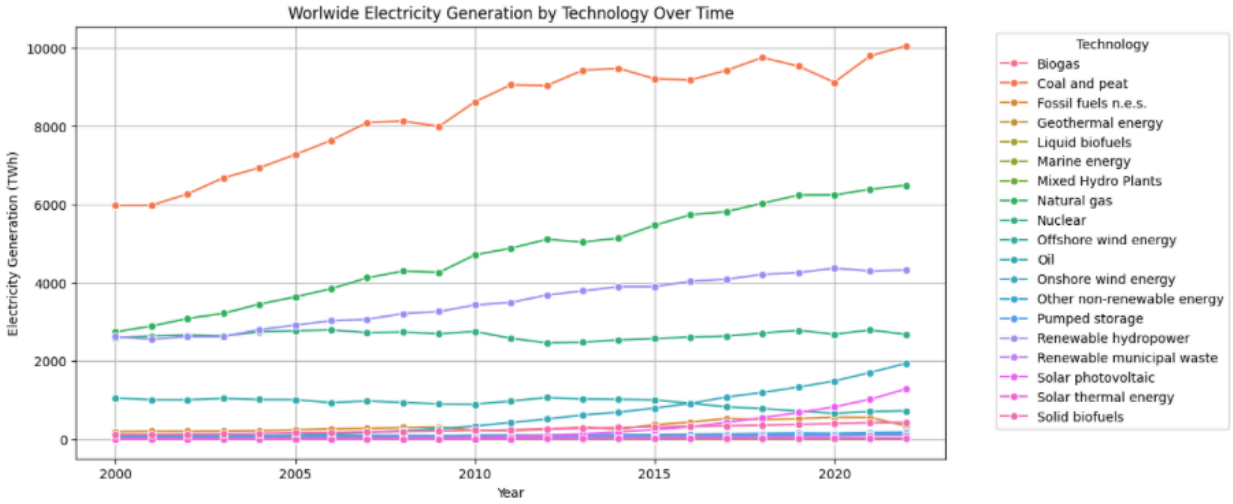

Figure: IRENA Worldwide Electricity Generation by Technology

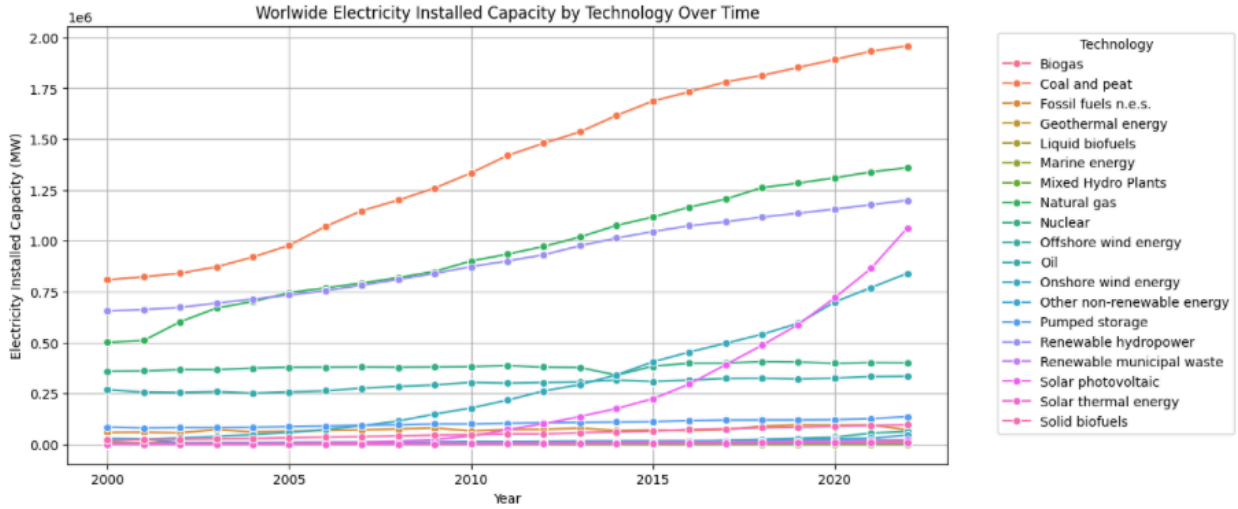

Figure: IRENA Worldwide Electricity Installed Capacity by Technology

As expected, electricity installed capacity increases in line with electricity generation to accommodate increased production. Some notable observations are that while coal is still the dominant leader in electricity generation worldwide, there is also rapid acceleration in natural gas and renewable hydropower energy generation. In addition, individual countries can be analyzed to view not only their energy production over time, but also any diversification or

sustainability efforts over time. Below, the energy generation graphs for Saudi Arabia and Brazil also from the IRENA dataset can be seen, grouped by type of technology.

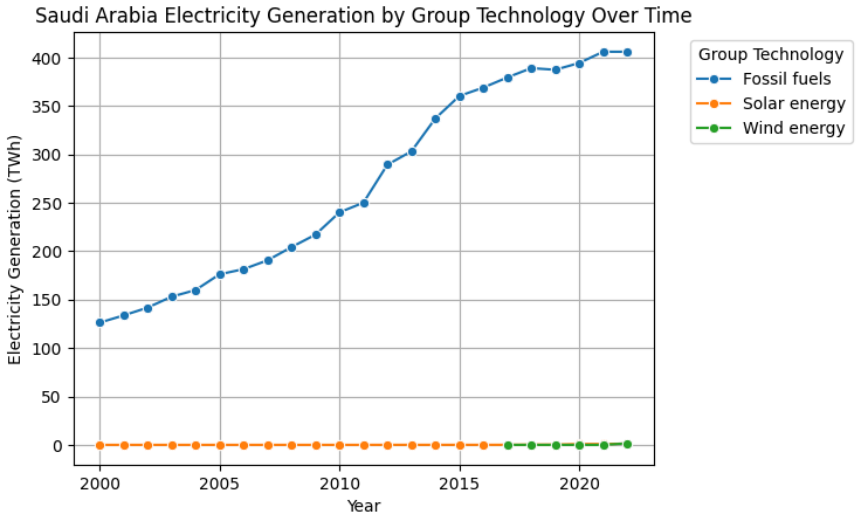

Figure: IRENA Saudi Arabia Electricity Generation

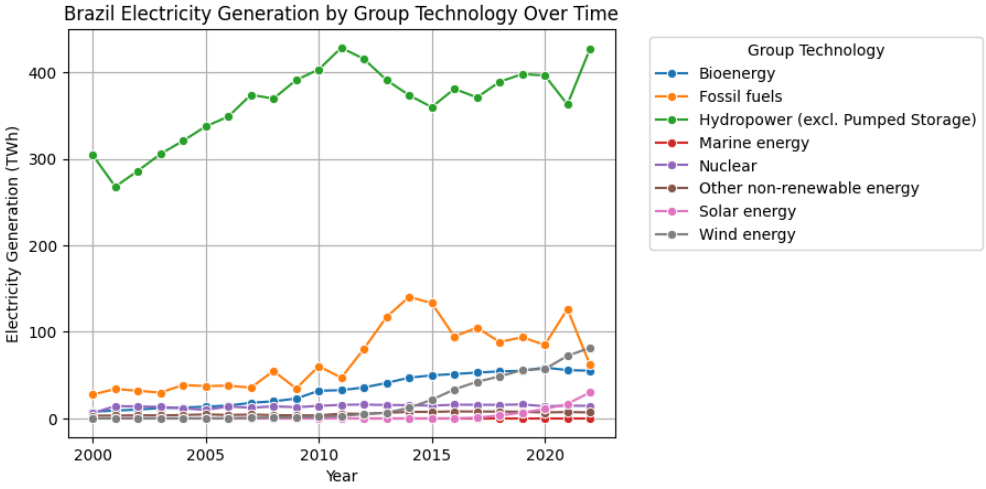

Figure: IRENA Brazil Electricity Generation

Since Saudi Arabia is the world's leading crude oil exporter, it's no surprise that they have little energy diversification, since their natural advantage lies heavily in their oil reserves. In contrast, somewhere like Brazil is a leader in hydropower electricity generation due to their elevation changes, large rivers, and high precipitation levels. In addition, their fossil fuel usage has seen a decrease in recent years while their wind energy generation has seen a significant uptick.

Therefore, it stands to reason that developing countries or even sub-regions with similar natural resources can look towards Brazil as an example of successfully leveraging their environment to work towards sustainability. In the case of Saudi Arabia, while it may not be feasible to steer entirely away from their reliance on oil as an export and resource for electricity generation, they can look towards their peers who excel in solar and wind energy generation so that their existing small production numbers in those categories can meaningfully scale.

Shifting perspective to the U.S. energy market through the dataset sourced from the U.S. Energy Information Administration (Morgado, 2023), energy trends can be analyzed on a state and monthly level. This dataset follows the same format as the previous two that were discussed, with energy generation broken down in terms of energy source, but it also contains monthly data which adds another level of nuance. For example, energy generation in relation to months can be seen below, plotted as boxplots.

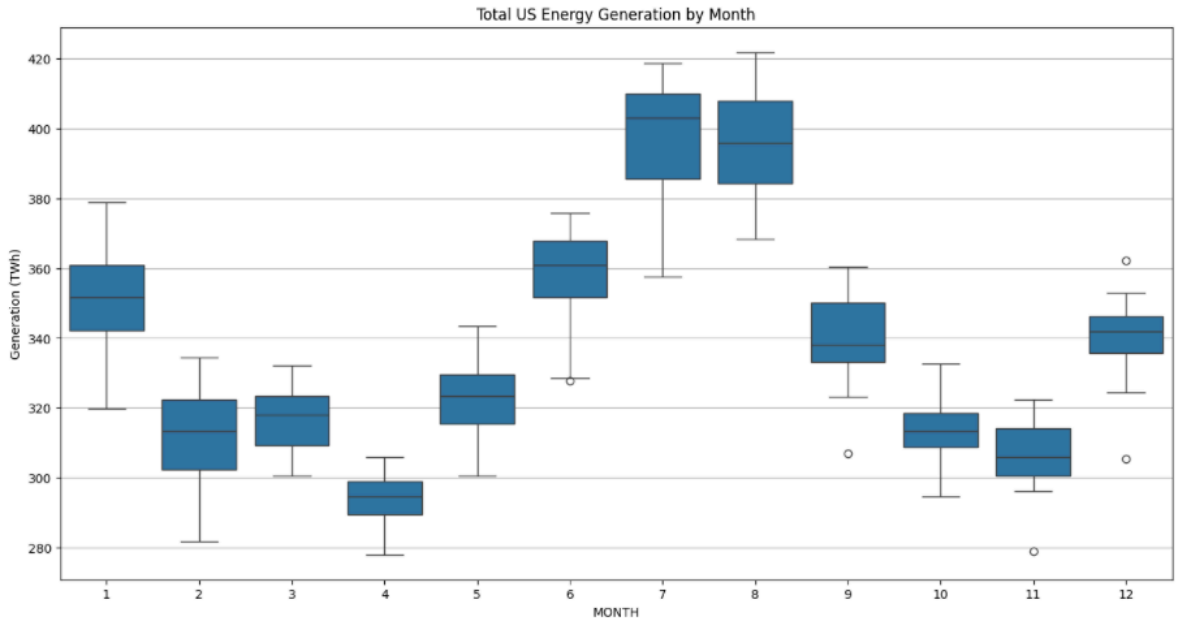

Figure: U.S. Energy Generation 2001-2022 Monthly Data

This monthly data, aggregated over the span of 20 years, shows total U.S. energy generation through the months. As is easily seen, it is quite cyclical, with energy generation peaking in the

summer months, which is likely due to increased demand for air conditioning and other cooling applications. Although this seems intuitive, visualizing quantitative values and distribution spreads of this data aggregated over the last two decades can identify trends, which can help energy suppliers and grid operators plan future electric grid output to meet the necessary demand. This can balance the supply-demand relationship and ensure a reliable power supply during these times.

A final exploratory visualization to highlight can be seen with the secondary global dataset (Hossainds, 2023). As a reminder, this dataset mainly split global renewable energy production trends by focusing solely on energy generated by geo biomass, solar, wind, and hydro. From a bivariate point of view, these different modes of production can be visualized with an area plot to demonstrate the dominant renewable energy generation methods as well as its growth due to global clean energy initiatives.

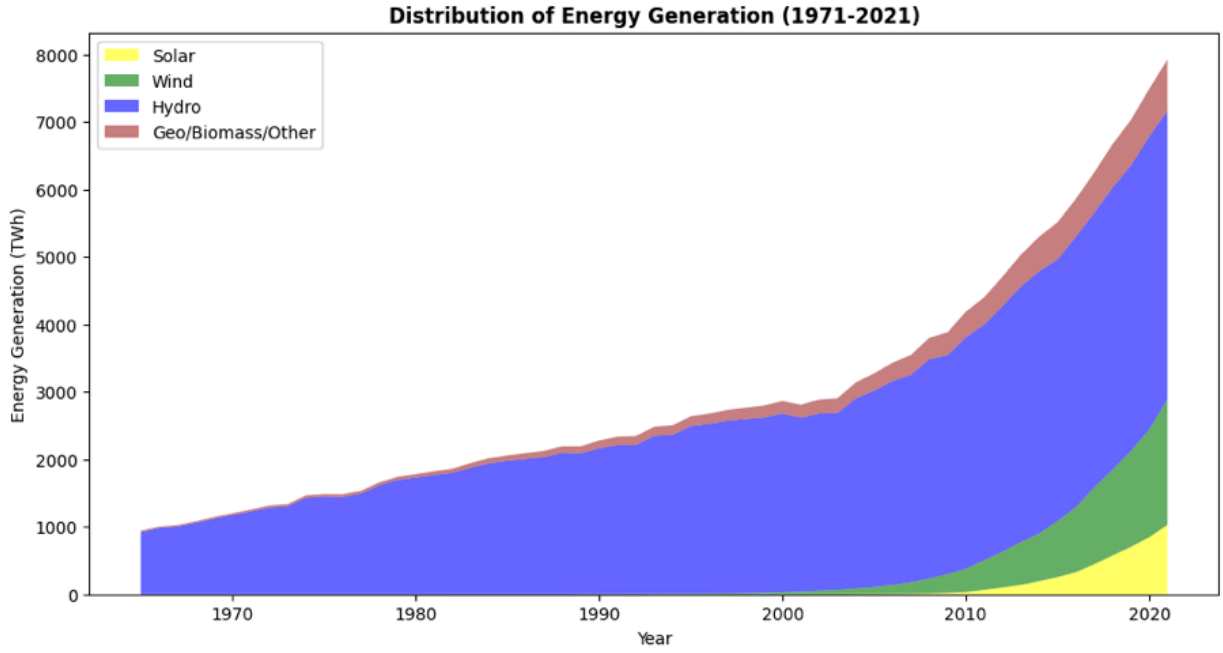

Figure: Renewable Energy Worldwide Area Plot

This area plot provides numerous insightful observations into renewable energy trends. From 1970 to 2000, there is a steady adoption of renewable energy worldwide, but hydro generation was essentially the sole source of renewable energy for those three decades with very slight contribution of geo/biomass/other generation. However, at the turn of the century, global initiatives focused on clean energy adoption began to take effect in a dramatic fashion. There is a clear significant increase in rate of energy generation at this point, with wind and solar also being added into the mix of methodologies. While wind and solar contributions to global renewable energy production are still dwarfed by hydro, they are increasing at an exponential rate, especially in just the past decade. Overall, the trend illustrated by this area plot is encouraging for policy makers and supporters of renewable energy adoption as it provides quantitative evidence that the shift in production is apparent and substantive compared to previous decades. While the world is still far from a full transition to renewable energy, it justifies that a clean energy transition is possible with an impressive rate of growth and new alternative technologies, such as hydrogen or nuclear, may even significantly add to this chart in the years to come just as wind and solar did in the previous two decades.

Section 4: Conclusions

In conclusion, all three datasets provided specific nuanced insights into energy trends both worldwide and on a regional/state level, complementing each other well. While two of the datasets focus on a global scale, one dataset (Har-Var, 2024) is inclusive of non-renewable energy which enables direct comparisons with renewable energy, and the other global dataset (Hossainds, 2023) is solely focused on renewable energy and has data that spans back to the 1960s. An aspect of the datasets that was unexpected was the formatting. Since all the data for each dataset was combined in a sort of long list with constant repeating series for years and types

of production/technologies for each region/country/state, the initial distributions of the numeric data were skewed and misleading. Only after grouping the data into specific subsets during bivariate analysis did the data make sense in a logical way. Relatedly, an aspect of the datasets that was expected was the relationship between (renewable) energy generation and time. Since the technological boom at the turn of the century and even present day with the energy demands of artificial intelligence, a positive correlation was expected. As populations increased, as technology adoption increased, and as renewable energy initiatives were being enacted globally, energy output needed to dramatically increase in recent decades to meet the demand. The specifics of that positive relationship were interesting to see, as indicated by the exploratory data analysis of the three datasets. The change in rate of renewable energy adoption, the cyclical nature of U.S. energy generation, and global renewable energy trends with respect to traditional fossil fuels were all intriguing observations.

No datasets were disqualified due to significant issues as they all were sufficiently cleaned for the most part, with just some reorganizing and grouping need for logical analysis. By this analysis, a clear supervised learning model to further explore would be regression, in order to predict expected energy output and demand on both a global and regional level. As noted with the U.S. monthly data, accurately predicting energy demands would allow grid operators to ensure stable and reliable power. An unsupervised analysis would be clustering, where geographically similar countries could be grouped together to determine most efficient pathways to renewable energy adoption. As seen in the IRENA dataset Brazil example, leveraging natural resources is important to unlocking renewable energy potential. Clustering countries with similar renewable energy outputs can lead to broader discussions of their successful adoptions. Combined with additional data perhaps on environmental and geographical features of countries,

underdeveloped countries can be targeted as candidates for renewable energy production based on their similarities to current leaders.

The questions and concerns posed by the creators and sources of the datasets include analyzing current trends for energy research, assisting policymakers in understanding the energy landscape, and how rapidly is the production of renewable energy changing with respect to energy mix? This exploratory data analysis was a good first step in answering these concerns through the previous visualizations and explanations. The trends on a global scale for energy generation/demand are clearly rising, and a more nuanced perspective could be examined on a country or regional scale. Policymakers would benefit from further analysis to see how feasible the shift to sustainable energy is in their respective regions or states and how rapidly production can ramp up with correct incentives and initiatives in place. Finally, energy mix was seen to be diversifying and production is significantly scaling, with possible new technologies on the horizon. All in all, this initial analysis accurately addresses these concerns or incentivizes further analysis into these concerns.

References

Har-Var (2024). *Global Renewable Energy Production (2000-2022)* [Data set]. Kaggle. https://www.kaggle.com/datasets/shakaal/global-renewable-energy-production-2000-2022/data

Hossainds, B. (2023). *Renewable Energy World Wide: 1965~2022: 02 modern-renewable-energy-consumption* [Data set]. Kaggle. https://www.kaggle.com/datasets/belayethossainds/renewable-energy-world-wide-19652022?select=02+modern-renewable-energy-consumption.csv

Morgado, K. (2023). *US Energy Generation 2001-2022* [Data set]. Kaggle. https://www.kaggle.com/datasets/kevinmorgado/us-energy-generation-2001-2022