Renewable Energy Data Analysis

Kevin Carney

Fatina Hamadi

Chris Hicks

Kade Rivers

UTSA Data Analysis Bootcamp

Project 1 Group Projects

Professor Booth

June 17, 2024

Abstract

This data science project is based on an educational dataset from Kaggle with randomized data for renewable energy systems with the intent to analyze the data, produce visualizations, and present on the findings to help investors make informed decisions using data analytics. The data science project was done by four UTSA Data Analysis Bootcamp students. The task was to use Python, Jupyter Notebook, and Matplotlib to visualize a dataset for a data science project. The research examined renewable energy parameters such as installed capacity, energy production and consumption, storage efficiency, financial investments, and environmental implications, and job creation.

Methodology

Our analysis utilized a Kaggle dataset on renewable energy systems to analyze and create visualizations and a presentation for a potential investor. As a team of four, we used Jupyter Notebook, Python, and Matplotlib to analyze and visualize data.

Data Cleaning and Preprocessing

We loaded the dataset into a Pandas DataFrame, examined the structure, identified constraints, and fixed discrepancies. We removed rows and supplied missing values using appropriate methods to improve understanding. We categorized numerical codes reflecting renewable energy kinds and other category characteristics.

Data Grouping and Aggregation Analysis

The data was categorized by renewable energy type to determine total and average metrics such as installed capacity, energy output, storage efficiency, financial investments, and environmental implications.

Composite scoring, normalization

We normalized the variables and calculated a composite score to compare renewable energy sources fairly. The performance of each energy source was evaluated using many parameters to create a composite score.

Visualization

Matplotlib was used to create visuals that revealed renewable energy performance. Bar charts helped us understand each energy source's capabilities.

Teamwork

Our team worked through paired programming sessions, communication through various sources, and further research into viable datasets throughout this project has increased our awareness and abilities as individuals and as team members. We met often to review progress, share discoveries, and integrate individual contributions into the study. Git managed code versions and facilitated cooperation. Continuing to build on our knowledge gained from this educational dataset, we can recreate and modify our analysis and visualization skills with scientifically accurate datasets and research to share accurate data with future companies and investors.

Analysis

Question 1, Kevin Carney

What kind of renewable energy sources have the highest energy output compared to their energy consumption?

As we began thinking about the dataset before us, one of the critical questions we sought to answer was which energy type was the most effective. We started this process by comparing the energy production vs. consumption for each type of energy.

Our initial attempts at visualizations for these relationships involved creating scatter plots based on these factors, with further refinements to include scaling the size of the pips based on storage efficiency. These visualizations ended up being inconclusive, due to the amount of data points in our data set. Eventually we settled on representing this relationship via a bar graph. The data was aggregated based on energy types, and an additional column was added to show the difference between production and consumption. The bars were then overlaid on top of each other to illustrate the difference between these two characteristics for each energy source. As each of the energy types were shown to have surpluses, it was simply a matter of discerning which had the highest surplus. Using these results, a horizontal bar graph was then generated to more clearly show the relative difference, and rank each energy type accordingly.

Fig 1 *- Energy Production vs Consumption Bar Graph (Code)*

```
y = bar_df["Energy_Production_MWh"]
y2 = bar_df['Energy_Consumption_MWh']
x = bar_df["renewable_energy_source"]


# Step 2: Create the Canvas
plt.figure(figsize=(8,6))

# Step 3: Create the skeleton plot
plt.bar(x, y, color="#46FD5F", label='Energy Production')
plt.bar(x, y2, color= "#982121", label='Energy Consumption')

# Step 4: Customization
plt.xlabel("Type of Renewable Energy", fontsize=12, fontstyle="italic")
plt.ylabel("Energy Produced / Consumed (MW/h)", fontsize=12, fontstyle="italic")
plt.title("Energy Production vs Consumption", fontsize=16, fontweight="bold")
plt.ylim(100000, 300000)

plt.grid(axis="x", color="lightgrey", linestyle="--", alpha=0.5)
plt.xticks(rotation=45)

# Step 5: Show/Save
plt.show()
```

Fig 2 *- Energy Production vs Consumption Bar Graph (Visualization)*

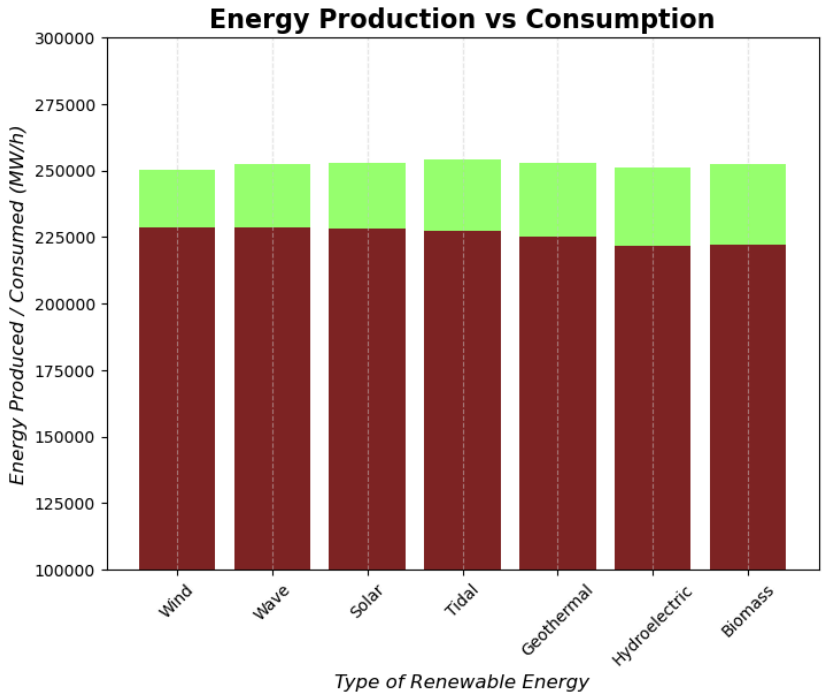
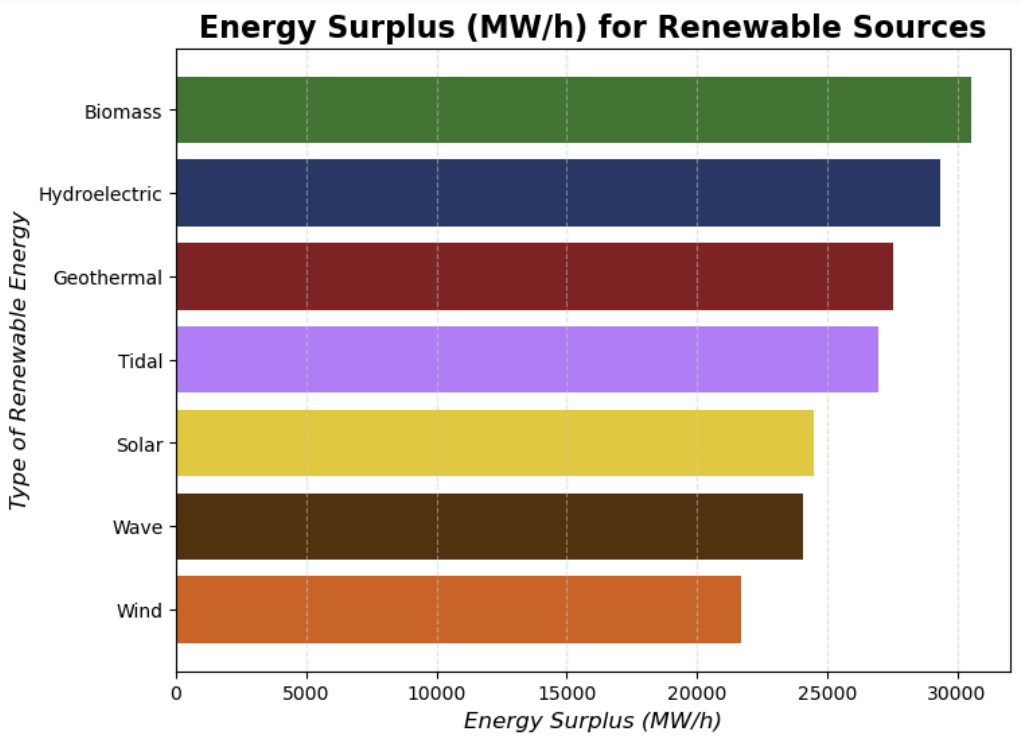


Fig 3 *- Energy Surplus - Horizontal Bar Graph (Code)*

```
y = bar_df["Energy_Surplus_MWh"]
x = bar_df["renewable_energy_source"]

c_colors = ['#E66012', '#5F2F08','#F5C31E', '#B789FA', '#982121', '#173F66', '#23732D']
# Step 2: Create the Canvas
plt.figure(figsize=(8,6))

# Step 3: Create the skeleton plot
plt.barh(x, y, color=c_colors)

# Step 4: Customization
plt.xlabel("Energy Surplus (MW/h)", fontsize=12, fontstyle="italic")
plt.ylabel("Type of Renewable Energy", fontsize=12, fontstyle="italic")
plt.title("Energy Surplus (MW/h) for Renewable Sources", fontsize=16, fontweight="bold")

plt.grid(axis="x", color="lightgrey", linestyle="--", alpha=0.5)

# Step 5: Show/Save
plt.show()
```

Fig 4 - *Energy Surplus - Horizontal Bar Graph (Visualization)*



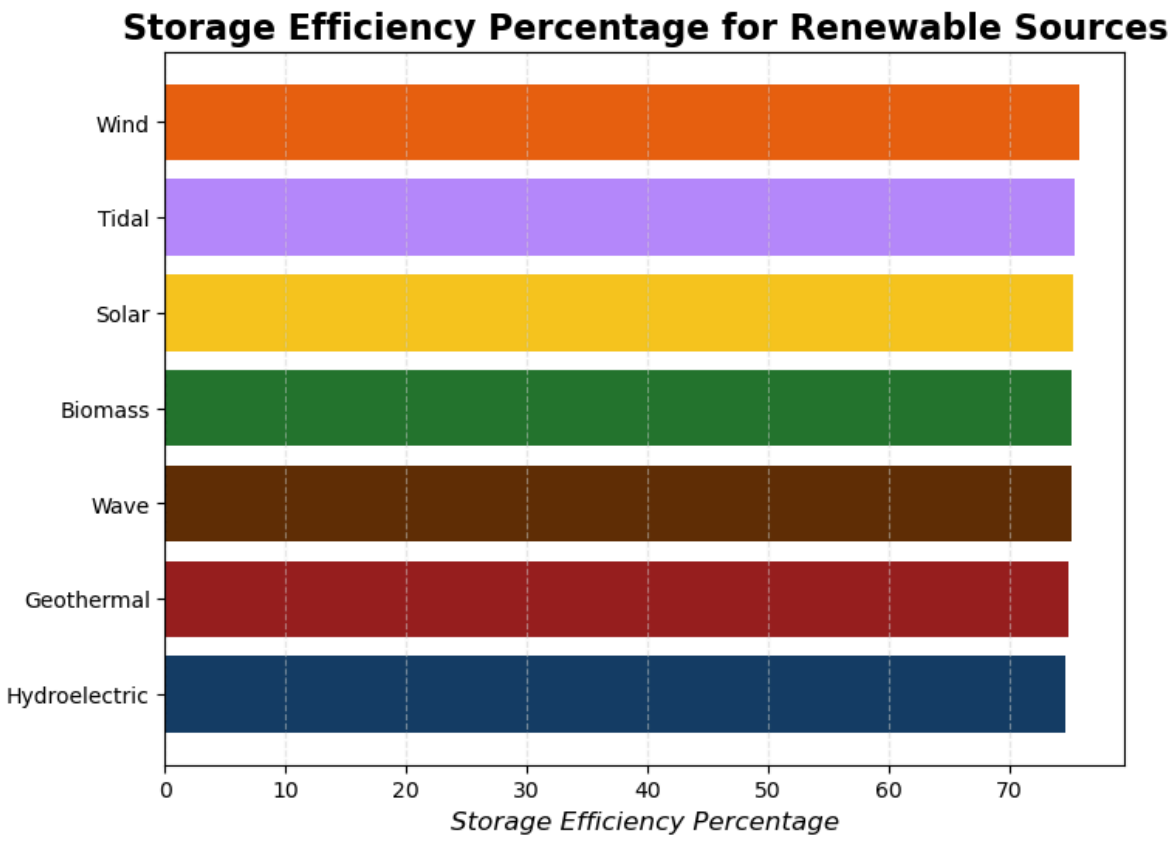
Based on our data, the energy type with the highest energy surplus was Biomass, followed by Hydroelectric and Geothermal. We found this surprising because the most widespread renewable energy sources (Solar and Wind) also had the lowest energy surplus based on the data in this dataset.

Wanting to still pursue storage efficiency as a way to understand additional factors that could be making Solar and Wind more popular outside of this dataset, we developed additional bar graphs that reviewed the relationship between energy type and storage efficiency. Once again we discovered an odd aspect of this dataset, in terms of mean storage efficiency each energy type had less than 1% variation in comparison to the others. Wanting to understand why, we dug into the data a little bit more and determined that the reason for this peculiarity is based on the fact that this is a series of projections based on potential energy sources as opposed to recorded outputs from in use systems.

Fig 5 - *Storage Efficiency Percentage (Code)*

```
storage_bar_df = df.groupby('renewable_energy_source')[['Energy_Production_MWh', 'Storage_Efficiency_Percentage']].mean().reset_index()
# bar_df["Energy_Surplus_MWh"] = bar_df.Energy_Production_MWh - bar_df.Energy_Consumption_MWh

storage_bar_df = storage_bar_df.sort_values(by = "Storage_Efficiency_Percentage", ascending = True)

k_colors = ['#173F66', '#982121','#5F2F08', '#23732D', '#F5C31E', '#B789FA', '#E66012']


y = storage_bar_df["Storage_Efficiency_Percentage"]
x = storage_bar_df["renewable_energy_source"]

display(storage_bar_df)

# Step 2: Create the Canvas
plt.figure(figsize=(8,6))

# Step 3: Create the skeleton plot
plt.barh(x, y, color=k_colors)

# Step 4: Customization
plt.xlabel("Storage Efficiency Percentage", fontsize=12, fontstyle="italic")
plt.title("Type of Renewable Energy", fontsize=16, fontweight="bold")
plt.title("Storage Efficiency Percentage for Renewable Sources", fontsize=16, fontweight="bold")

plt.grid(axis="x", color="lightgrey", linestyle="--", alpha=0.5)

# Step 5: Show/Save
plt.show()
```

Fig 6 - Storage Efficiency Percentage (Visualization)

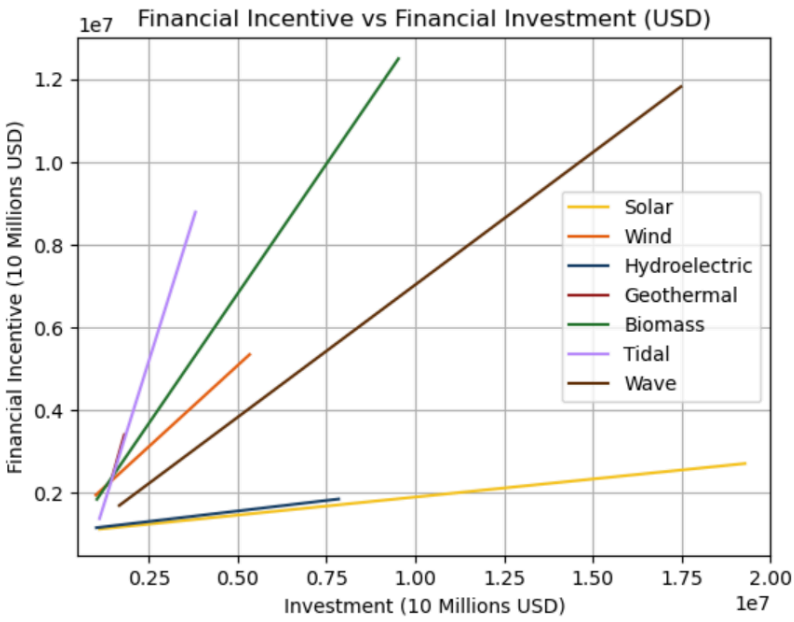


Now that we understood the nature of this relationship and data set, we felt that analysis of similar data sets could help us answer some additional questions. Part of the purpose of renewable energy is the reduction of Air Pollution / Greenhouse Gases in support of the planet and public health. Though Biomass may have the highest energy surplus, per the U.S. Energy Information Administration, it is net neutral in terms of a carbon footprint. Conversely Solar and Wind Energy are shown to have little to no pollution because of their use. Furthermore, depending on the type of biomass being used, it can take between 3 to 4 weeks in the case of microalgae, or 3 to 5 years in the form of wood crops to be able to produce a product that can be viable for energy production. Solar and Wind by comparison are truly renewable resources that are available across most parts of the world every day. Solar sources alone can produce 173,000 terawatts of energy every day which, according to the Department of Energy, is 10,000 times the world's total energy use.

Despite the applicability of these renewable energy sources, there is another side to this topic that is worth giving consideration to, and that is the human or financial side of renewable energy. This subject was the focus of our next question.

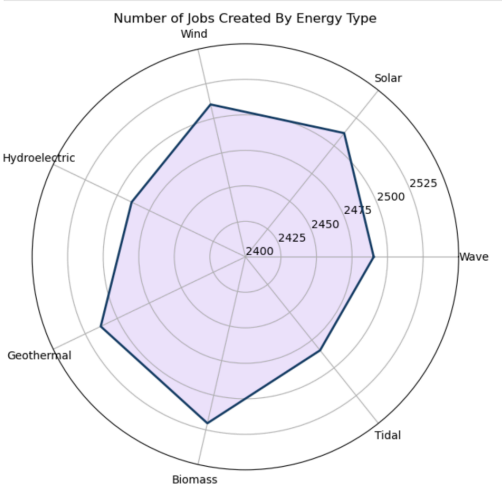
Question 2, Kade Rivers

According to this dataset, which renewable energy sources have the greatest financial incentives versus initial investment?



When digging into the data on financial investment versus financial incentive, we uncover our first data that displays some actual distinctions to help determine which type of renewable energy is the most likely to offer the best financial incentives. While there wasn't a time element within the dataset, there is a clear visualization from the dataset that both Tidal and Geothermal renewable energies will yield a higher incentive with a smaller investment. Renewable sources such as Solar and Hydroelectric energies require a higher investment amount with far less financial incentive, and Biomass, Wind, and Wave energies trend having a higher incentive as the investment increases.

Although financial incentive is a top priority for many companies, a complete analysis of the data along with additional resources is key to making the best informed decision. Another indicator for a company to make a decision on an investment would be the number of potential jobs created from the investment. The data within this radar chart indicates that the type of energy invested in will have a fairly minimal impact on job creation. However, an investor or company may make note of the amount of financial incentive from the above line graph, determine the potential jobs created from the initial investment, and populate speculated data on how higher financial incentives would, in turn, lead to future investments and more jobs created over time.



While this dataset cannot help inform an investment decision, data analysts and companies need to have multiple datasets and resources when making such large decisions. Through our analysis and additional research, we were able to determine that this dataset was not scientifically accurate. Further digging into the dataset led us to the discovery that this dataset is for educational purposes only and should not be used to make any sort of financial investment. The dataset's limitations were made apparent in our final question about which renewable energy is the most popular.

Question 3, Chris Hicks

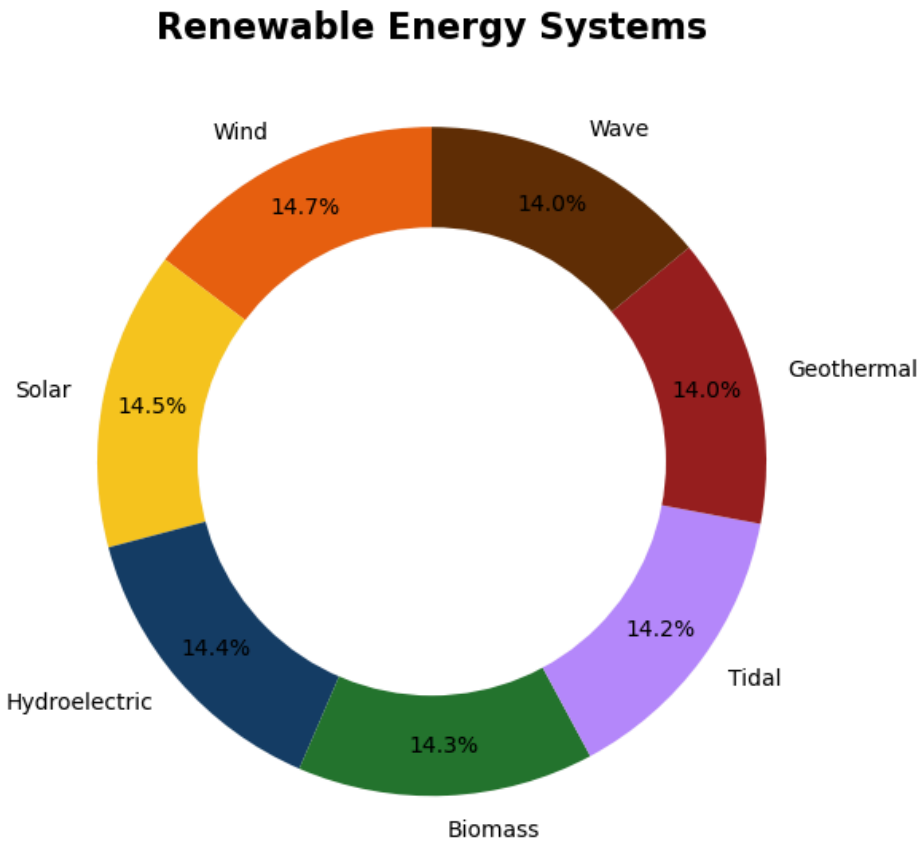
Which of the several forms of renewable energy is the most widely used?

While looking at the dataset to determine the most widely used renewable energy source we looked into prioritizing job creation, environmental cleanliness, cost-effectiveness, and efficiency for the analysis. We created visual representations, like the donut plot, only to find that the top three contenders (solar, wind, and hydro) were closely matched in every aspect.

Fig 9 - *Donut Chart (code)*

```
data = df_value["count"]
labels = df_value["renewable_energy_source"]
mapped_colors = [colors[label] for label in labels]

# Step 2: Make the canvas
plt.figure(figsize=(8, 6))

# Step 3: Make the plot
plt.pie(data, labels=labels, colors=mapped_colors,
        autopct="%1.1f%%", shadow=False, startangle=90, pctdistance=0.85)

# Step 3.5: Add in the donut hole
# draw circle
center_circle = plt.Circle((0, 0), 0.70, fc='white')
fig = plt.gcf()

# Adding Circle in Pie chart
fig.gca().add_artist(center_circle)

# Step 4: Customizations
plt.title(f"Renewable Energy Systems\n", fontweight="bold", fontsize=16)
plt.axis("equal")

# Step 5: Save/Show
plt.show()
```

Fig 10 - Donut Chart (Visualization)



Upon analyzing our findings, there was a distinct lack of numerical and visual differences. This led to additional research on renewable energy sources. Online sources with scientific data and additional studies led us to the discovery that solar energy created the most jobs globally (12.7 million). Solar energy was also the cleanest, and required minimal maintenance. This revelation led us to question the validity of our dataset. In an attempt to introduce some differentiation, we created new columns and ratios, but ultimately, we realized that our dataset was curated and needed to definitively indicate which energy source was superior, regardless of the criteria used for comparison.

For further research and study on this topic, we have included references that helped confirm our suspicions of an inaccurate dataset.

Reference: Which Renewable Energy Sources Are Most Reliable?

Reference: Which renewable energy created the most jobs?

Linear Regression, Fatina Hamadi

Wind Energy Relationships

This project focuses on utilizing a linear regression model to analyze various parameters related to renewable energy production. Having a clear understanding of these aspects is crucial for stakeholders to enhance the efficiency and production of renewable energy projects. The dataset provides comprehensive information on various aspects of renewable energy, including sources, capacity, production and consumption, storage, grid integration, investment, funding, incentives, emission reduction, pollution reduction, and job creation. The linear regression model offered valuable insights into the connections between various features and the energy production of renewable energy systems. The model's performance was assessed by evaluating various metrics, including the slope, intercept, r-value, p-value, and stderr. Next, the regression lines are plotted on the scatter plot to visualize the actual versus predicted values.

These visualizations include Greenhouse Gas Emissions vs. Air Pollution Reduction for wind energy, Greenhouse gas Emissions vs. initial Investment for wind energy, Air pollution reduction vs. initial Investment for Wind Energy, Air pollution Reduction vs. job Creation, and for Air pollution Reduction, Greenhouse gas Emissions VS Fully Integrated, Partially Integrated, Minimal Integration, Isolated Microgrid. The linear regression model offered valuable insights into the connections between various features and the energy production of renewable energy systems. The scatter plot effectively showcased the model's predictive capability and highlighted any discrepancies between the actual and predicted values.

Fig 11 - *Air Pollution Reduction vs Initial Investment*

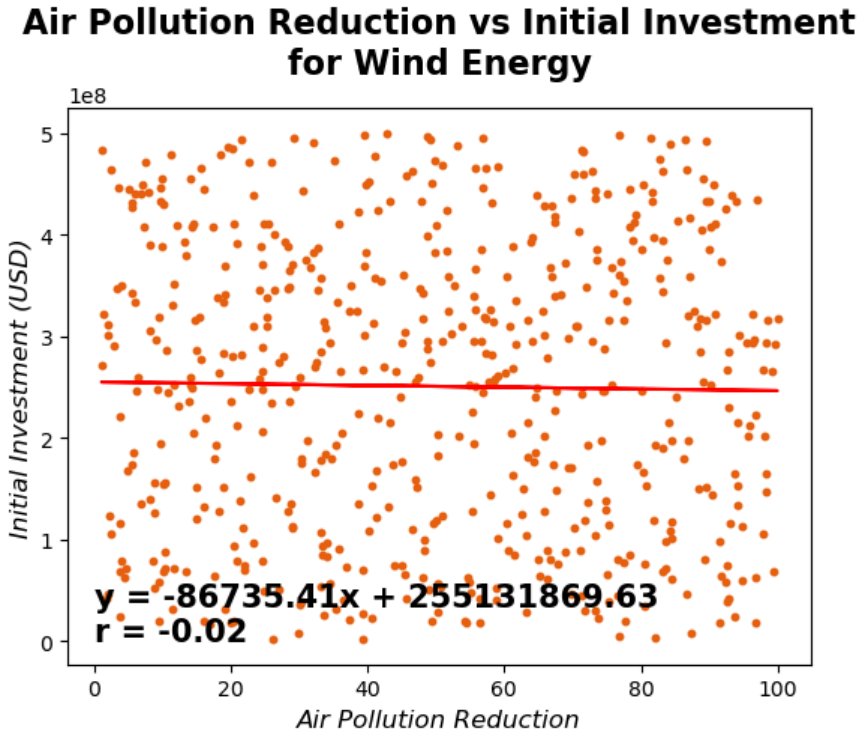
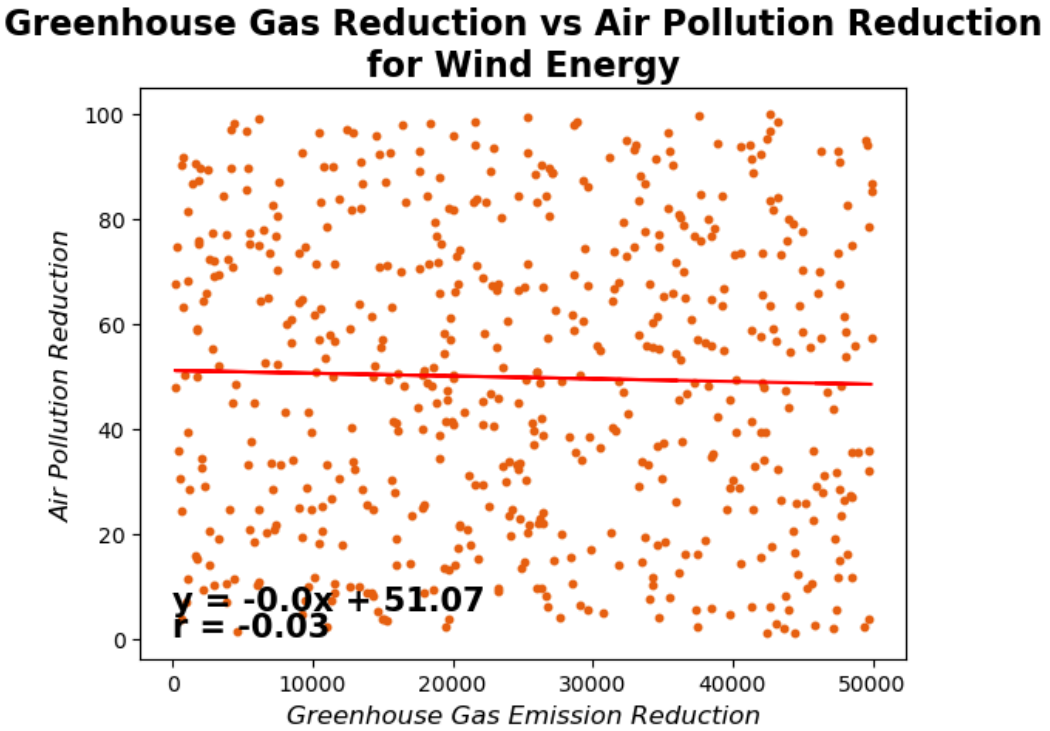


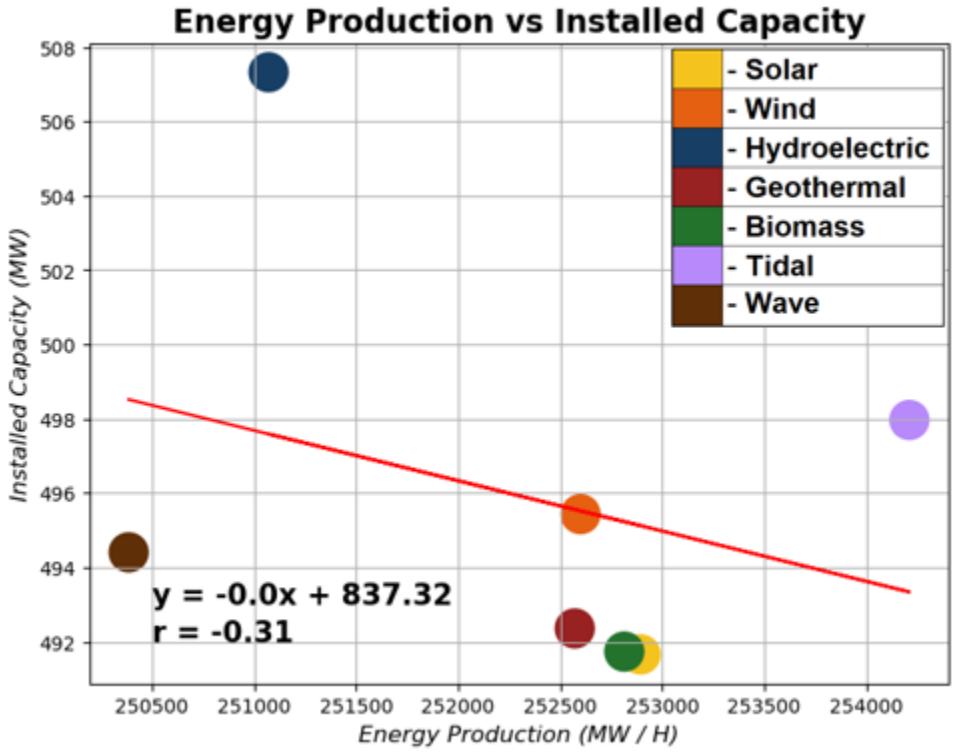
Fig 12 *- Greenhouse Gas Reduction vs Air Pollution Reduction*



Here are the main findings from the linear regression analysis:

There is a strong negative correlation between installed capacity and energy production, suggesting that larger capacity installations tend to generate less energy.

Fig 13 - Energy Production vs Capacity



The success of renewable energy projects heavily relied on financial factors such as initial investment and financial incentives. The higher the investments and incentives, the greater the energy production.

The environmental benefits of reducing GHG emissions and air pollution were found to be closely linked to energy production. This highlights the dual advantage of renewable energy projects, as they contribute to energy generation and play a crucial role in protecting the environment.

In conclusion

This study highlights the effectiveness of linear regression in analyzing and predicting the performance of renewable energy projects. By comprehending the various factors that impact

energy production, stakeholders can make well-informed decisions to maximize the effectiveness and efficiency of renewable energy systems. This analysis provides valuable insights that can inform planning and policy-making efforts to improve the adoption and effectiveness of renewable energy technologies.

Call to Action

As the world moves forward into a future that needs increasing amounts of energy to support its growing population. It is important that clean and affordable renewable energy alternatives are available to all citizens. This data set, while simulated, should lead to additional studies using real world, measurable data, in the hopes of working through the difficulties that renewable energy implementation currently faces.

Conclusion

This dataset was ultimately created for learning purposes in place of scientifically proven and accurate data that can be transferred for recommendations or speculations. However, even within a dataset of randomized numbers that didn't have the expected outcomes we anticipated based on our prior knowledge, we were still able to analyze and provide visualizations for this dataset. While this dataset concludes that all of the forms of renewable energy are fairly even for production, consumption, funding, and jobs created, we were able to determine the dataset wasn't factual through our analysis. Following that, we were able to find some additional resources further proving our assumptions and confirming that the dataset was randomized for educational purposes only.

Project Difficulties and Learning Opportunities

As a team of new data analysts on their first project, we encountered several difficulties that turned into great learning opportunities for all of us. The first, and biggest learning outcome was to dig deeper into the original source of the dataset to ensure the accuracy and viability of the data being analyzed. Along with sources for the original dataset, we learned why it is vital to have multiple datasets for a project and for a true analysis in order to compare and contrast the results for best predictions and accuracy.

The next learning opportunity came from the realization that the dataset lacked chronological and geographic information. This led to the inability for determined growth over time as well as advantages or disadvantages of geographic location that would arise for the various renewable energy types.

Through the challenges of our first project, we feel confident in the ability to recreate the details, analysis, and visualizations for future projects and future jobs we will have in the field of data analytics and visualization. Every obstacle is a chance for increased depth of knowledge and understanding and we look forward to our next opportunity.

Data Set Used

Link:
https://www.kaggle.com/datasets/girumwondemagegn/dataset-for-renewable-energy-systems
Creator: Girum Wondemagegn

References

Iea. “Executive Summary – Renewables 2023 – Analysis.” *IEA*, www.iea.org/reports/renewables-2023/executive-summary. Accessed 6 June 2024.

Reference: Which Renewable Energy Sources Are Most Reliable?
Reference: Which renewable energy created the most jobs?