PIMPRI CHINCHWAD EDUCATION TRUST's

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING

Report: RENEWABLE ENERGY DATA ANALYSIS

Subject Name: Statistical Data Analysis Using R

Subject Teacher Name: Dr. Nishi Gupta Ma'am

Presenter: Rouchi Dattatrey Mahajan

PRN: 123B1F134

Department: Information Technology (B)

Title: Renewable Energy Dataset Analysis

Link:

- Drive: https://drive.google.com/drive/folders/1o40y9KS9Td0ntGW8qExbP-b0qvxuFZOj?usp=sharing

- Kaggle: https://www.kaggle.com/datasets/girumwondemagegn/dataset-for-renewable-energy-systems

Description:

The Renewable Energy Dataset contains records of various renewable energy projects. It includes details about the types of renewable energy (solar, wind, hydroelectric, etc.), grid integration levels, installed capacity (in MW), energy production (in MWh), greenhouse gas emission reductions (in CO2), jobs created, initial investment, financial incentives, and air pollution index. The purpose of the dataset is to evaluate and compare the performance of different renewable energy sources across environmental, economic, and social parameters. This analysis helps to assess which energy sources offer the best overall benefit.

Last Updated:6 months ago

No of rows: 15001

No of columns:13

Problem Statements:

1.Relationship between initial investment and financial incentives:

Problem Statement: Investigate the relationship between initial investment requirements and available financial incentives across different renewable energy technologies. Are there significant variations in investment needs and support levels among different renewable energy sources, and which technologies demonstrate the most favourable investment-to-incentive ratios?

2. Relationship between energy production and consumption patterns:
Analyze the relationship between energy production and consumption patterns across different renewable energy sources. How do production

capabilities align with consumption demands for each renewable technology? Which sources show the most balanced or efficient production-consumption ratios?

3.Air pollution reduction capability

Problem Statement: Analyze the effectiveness of different renewable energy sources in reducing air pollution. How do various renewable technologies compare in their air pollution reduction capabilities? Which renewable sources demonstrate the highest potential for improving air quality?

4.Return on investment

Problem Statement: How do different renewable energy technologies compare in terms of their return on investment, and which technologies offer the most financially efficient energy production relative to their investment costs?

5.Best Renewable Energy (Parameters wise)

Problem Statement: Evaluate and compare different renewable energy technologies across multiple parameters (installed capacity, energy production, energy consumption, energy storage capacity, energy storage efficiency ,investment costs, financial incentives, GHG reduction, funding source and jobs created) to identify the most efficient and cost-effective renewable energy solutions.

Methodology

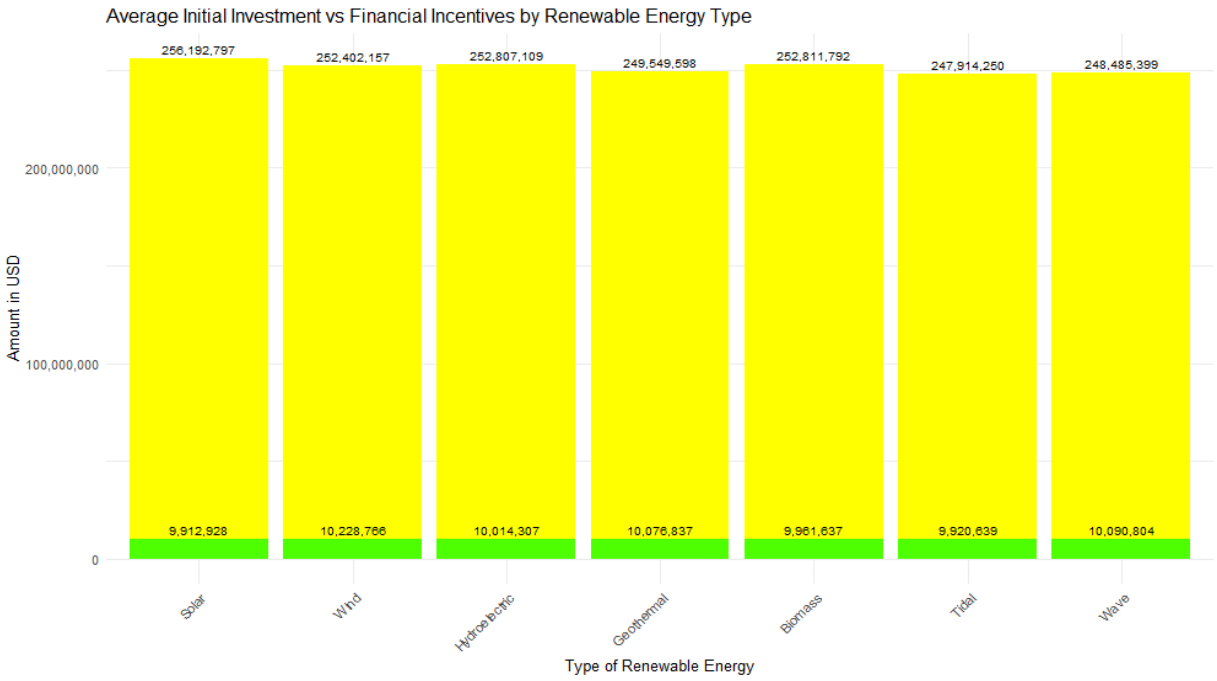
Prerequisites loading:

```r
# Load necessary libraries
library(ggplot2)
library(dplyr)
library(reshape2)
library(scales)

# Load the dataset
data <- read.csv("C:/Users/mfcs/Documents/FA_1/energy_dataset_.csv")

# Convert numerical codes to descriptive labels for readability
data$Type_of_Renewable_Energy <- factor(data$Type_of_Renewable_Energy,
                                        levels = 1:7,
                                        labels = c("Solar", "Wind", "Hydroelectric",
                                                   "Geothermal", "Biomass", "Tidal", "Wave"))

data$Grid_Integration_Level <- factor(data$Grid_Integration_Level,
                                      levels = 1:4,
                                      labels = c("Fully Integrated", "Partially Integrated",
                                                 "Minimal Integration", "Isolated Microgrid"))

data$Funding_Sources <- factor(data$Funding_Sources,
                               levels = 1:3,
                               labels = c("Government", "Private", "Public-Private Partnership"))
```

Loading the dataset and necessary libraries. Also as the dataset contains numerical codes for energy types, grid integration, and funding sources. Converting these codes into descriptive labels makes the data more interpretable and readable.

1.Relationship between initial investment and financial incentives:

```
# Calculate averages for Initial Investment and Financial Incentives
average_investment_incentives <- data %>%
  group_by(Type_of_Renewable_Energy) %>%
  summarise(Avg_Investment = mean(Initial_Investment_USD, na.rm = TRUE),
            Avg_Financial_Incentives = mean(Financial_Incentives_USD, na.rm = TRUE))

# Bar plot for Initial Investment and Financial Incentives with readable numbers
ggplot(average_investment_incentives, aes(x = Type_of_Renewable_Energy)) +
  geom_bar(aes(y = Avg_Investment), stat = "identity", fill = "yellow") +
  geom_text(aes(y = Avg_Investment, label = scales::comma(Avg_Investment)),
            vjust = -0.5, color = "black", size = 3) + # Labels for Avg_Investment
  geom_bar(aes(y = Avg_Financial_Incentives), stat = "identity", fill = "green", alpha = 0.7) +
  geom_text(aes(y = Avg_Financial_Incentives, label = scales::comma(Avg_Financial_Incentives)),
            vjust = -0.5, color = "black", size = 3) + # Labels for Avg_Financial_Incentives
  labs(title = "Average Initial Investment vs Financial Incentives by Renewable Energy Type",
       x = "Type of Renewable Energy", y = "Amount in USD") +
  scale_y_continuous(labels = scales::comma) + # Use comma format for y-axis
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Investment Overview

- Initial investments compared across seven renewable energy types
- Solar leads with highest investment at $256.2 million
- Wave energy has lowest requirement at $248.5 million
- Narrow investment spread indicates comparable capital needs

Financial Incentives Analysis

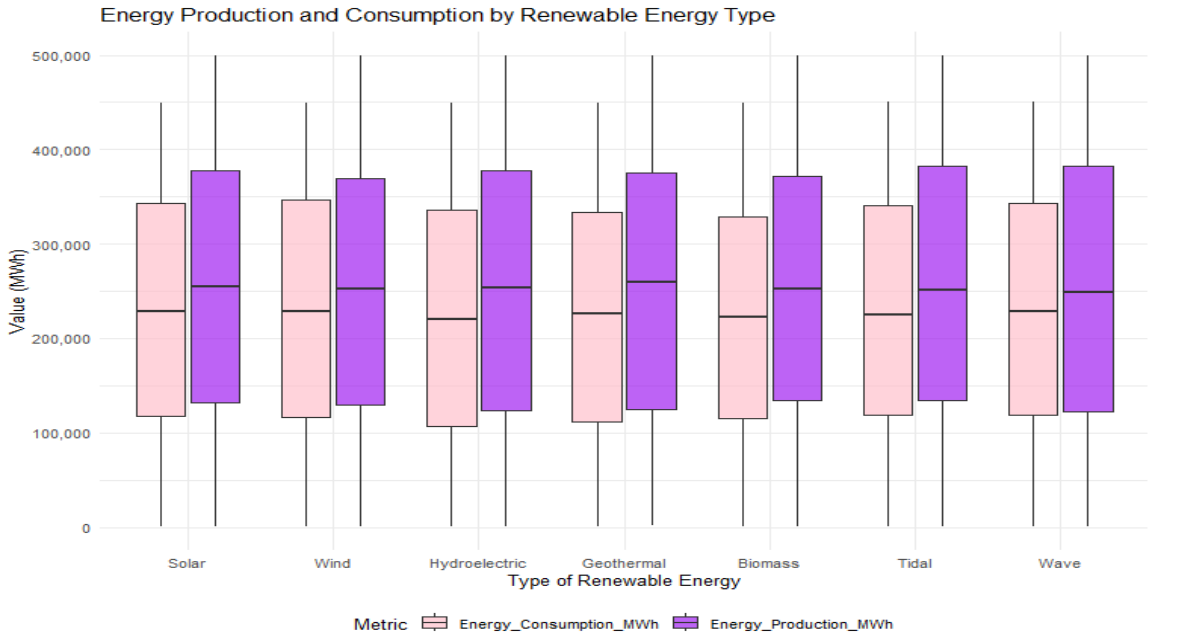
- Wind energy tops with $10.2 million in financial support
- Other technologies receive between $9.8-10.2 million
- Solar receives slightly lower incentives at $9.9 million
- Consistent 4% incentive-to-investment ratio across all types

Strategic Implications

- Uniform incentive distribution shows balanced policy approach
- Large investment-incentive gap indicates significant entry barriers
- Data benefits multiple stakeholders:
    - Policy makers
    - Investors
    - Project developers

2. Relationship between energy production and consumption:

```
# Reshape data for box plot
data_long <- data %>%
  select(Type_of_Renewable_Energy, Energy_Production_MWh, Energy_Consumption_MWh) %>%
  pivot_longer(cols = c(Energy_Production_MWh, Energy_Consumption_MWh),
               names_to = "Metric", values_to = "Value")

# Box plot for Energy Production vs Consumption with readable numbers
ggplot(data_long, aes(x = Type_of_Renewable_Energy, y = Value, fill = Metric)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Energy Production and Consumption by Renewable Energy Type",
       x = "Type of Renewable Energy", y = "Value (MWh)", fill = "Metric") +
  scale_fill_manual(values = c("Energy_Production_MWh" = "purple", "Energy_Consumption_MWh" = "pink")) +
  scale_y_continuous(labels = comma) + # Use comma format for y-axis
  theme_minimal() +
  theme(legend.position = "bottom")
```



Production Overview:

- Box plots show energy production vs consumption across seven renewables
- Production (purple) consistently higher than consumption (pink)
- Solar shows highest variability in production (largest box size)
- Wave energy demonstrates most stable production pattern (smallest box size)

Consumption Analysis:

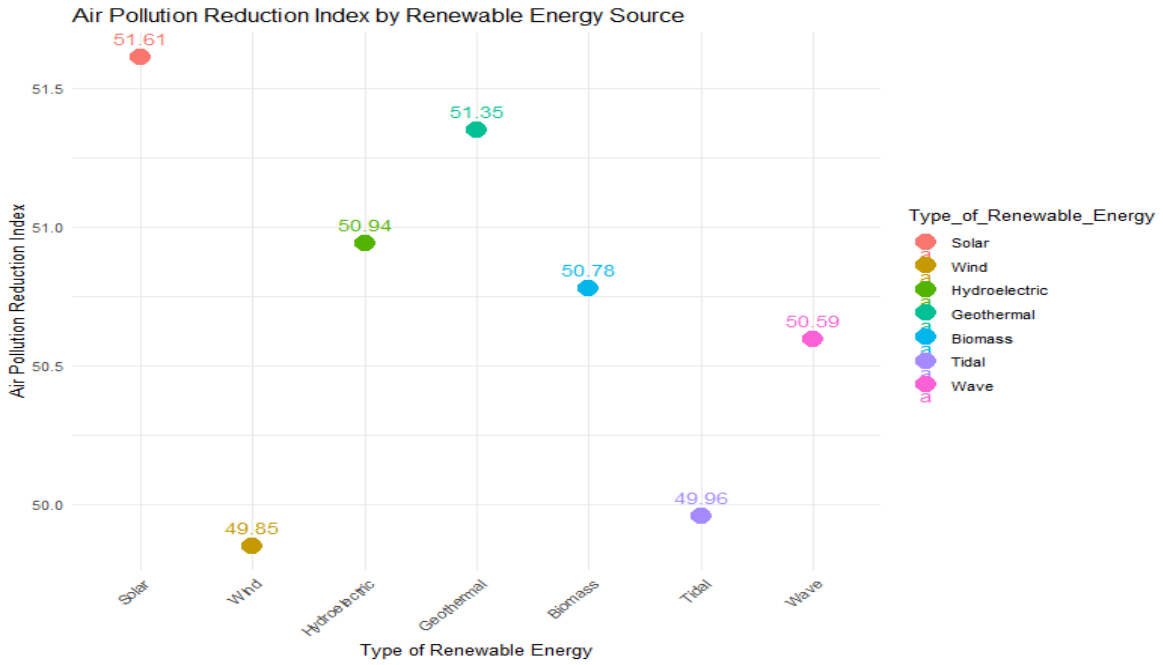
- Average consumption ranges between 200,000-250,000 MWh
- Consumption patterns relatively uniform across technologies
- Solar shows highest consumption variability
- Wave and Geothermal exhibit most stable consumption

Statistical Insights

- Median production exceeds median consumption for all types
- Interquartile ranges indicate production volatility varies by type
- Outliers minimal, suggesting reliable data distribution
- Production-consumption gap consistent across technologies

3.Air pollution reduction capability

```
# Scatter plot for Air Pollution Reduction Index with readable numbers
summary_data <- data %>%
  group_by(Type_of_Renewable_Energy) %>%
  summarise(Air_Pollution_Reduction = mean(Air_Pollution_Reduction_Index, na.rm = TRUE))

ggplot(summary_data, aes(x = Type_of_Renewable_Energy, y = Air_Pollution_Reduction, color = Type_of_Renewable_Energy)) +
  geom_point(size = 5) +
  geom_text(aes(label = round(Air_Pollution_Reduction, 2)), vjust = -1, size = 4) +
  labs(title = "Air Pollution Reduction Index by Renewable Energy Source",
       x = "Type of Renewable Energy", y = "Air Pollution Reduction Index") +
  scale_y_continuous(labels = comma) + # Ensure y-axis labels are readable
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Index Performance

- Solar leads with highest pollution reduction index at 51.61
- Geothermal follows closely at 51.35
- Wave energy shows lowest impact with 50.59
- Tidal energy has second-lowest rating at 49.96

Distribution Pattern

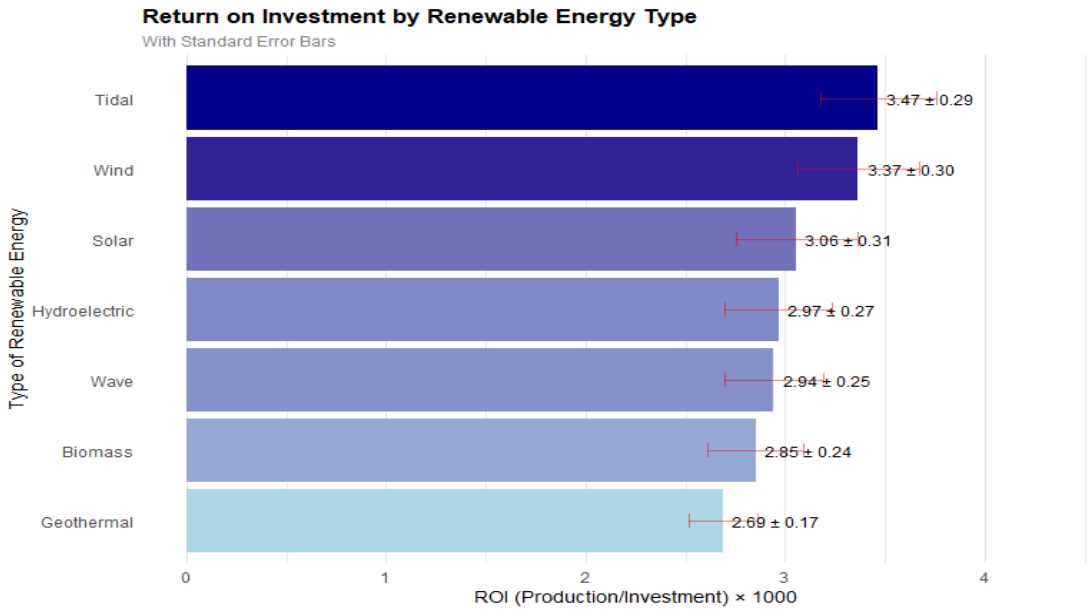
- Values range from 49.85 to 51.61 (narrow spread)
- Three distinct performance tiers visible:
    - Top tier: Solar, Geothermal (>51)
    - Mid tier: Hydroelectric, Biomass (50-51)
    - Lower tier: Wind, Tidal, Wave (<50)
- Overall spread of 1.76 points between highest and lowest

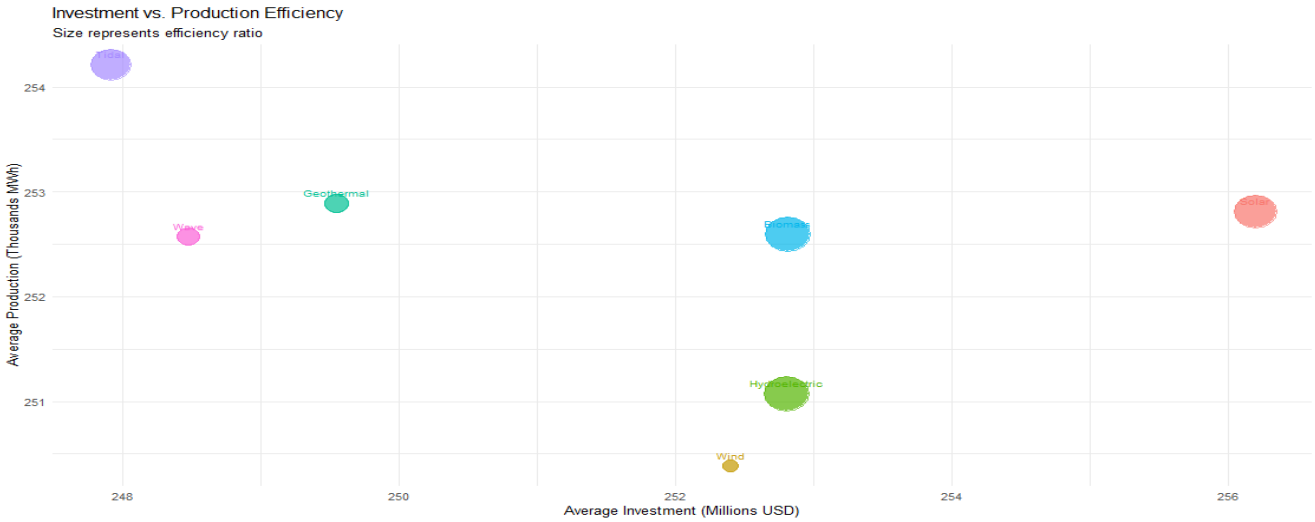
Comparative Analysis

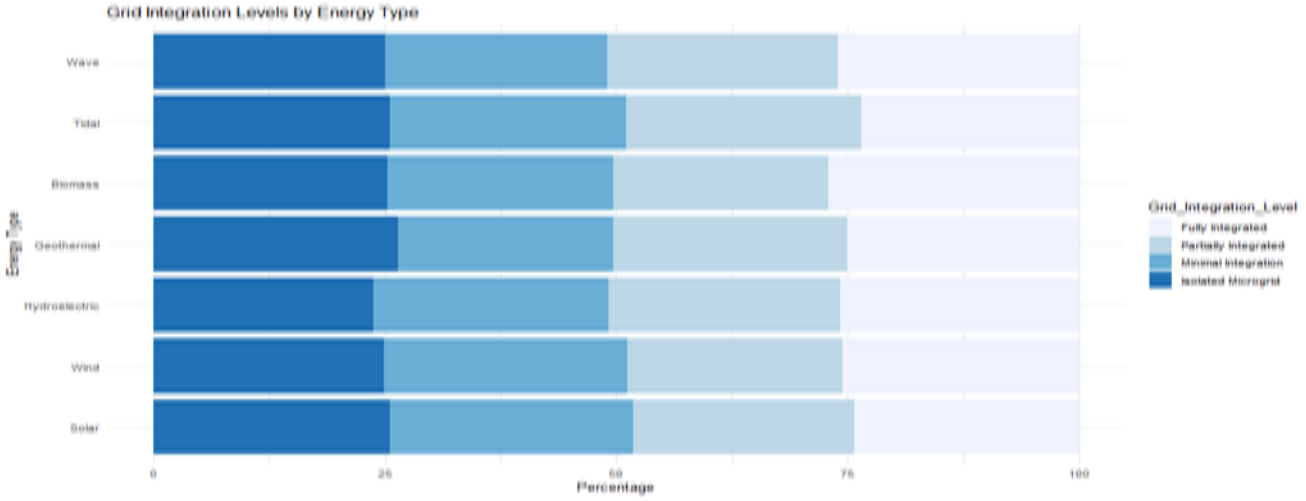
- Traditional renewables (Solar, Geothermal) show superior performance
- Marine-based solutions (Wave, Tidal) demonstrate lower reduction rates
- Wind shows surprisingly lower performance at 49.85
- Biomass maintains moderate effectiveness at 50.78

4.Return on investment

```r
# ROI Bar plot with error bars and readable numbers
ggplot(performance_metrics_with_error,
       aes(x = reorder(Type_of_Renewable_Energy, ROI_mean), y = ROI_mean)) +
  geom_bar(stat = "identity", aes(fill = ROI_mean)) +
  geom_errorbar(aes(ymin = ROI_mean - ROI_se, ymax = ROI_mean + ROI_se),
                width = 0.2, color = "red", alpha = 0.5) +
  geom_text(aes(label = sprintf("%.2f ± %.2f", ROI_mean, ROI_se)),
            hjust = -0.1, size = 3.5) +
  coord_flip() +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Return on Investment by Renewable Energy Type",
       subtitle = "With Standard Error Bars",
       x = "Type of Renewable Energy",
       y = "ROI (Production/Investment) x 1000") +
  scale_y_continuous(labels = comma) + # Format y-axis labels with commas
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(size = 10, color = "gray50"),
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 12),
    legend.position = "none",
    panel.grid.major.y = element_blank(),
    panel.grid.minor.y = element_blank(),
    panel.grid.major.x = element_line(color = "gray90"),
    plot.margin = margin(1, 2, 1, 1, "cm")
  ) +
  expand_limits(y = max(performance_metrics_with_error$ROI_mean +
                        performance_metrics_with_error$ROI_se) * 1.2)

# Comprehensive metrics by energy type
detailed_metrics <- data %>%
  group_by(Type_of_Renewable_Energy) %>%
  summarise(
    Avg_Investment = mean(Initial_Investment_USD, na.rm = TRUE),
    Avg_Production = mean(Energy_Production_MWh, na.rm = TRUE),
    Avg_Incentives = mean(Financial_Incentives_USD, na.rm = TRUE),
    Efficiency = mean(Energy_Production_MWh / Energy_Consumption_MWh, na.rm = TRUE),
    GHG_Reduction = mean(GHG_Emission_Reduction_tCO2e, na.rm = TRUE),
    Jobs_per_Investment = mean(Jobs_Created / Initial_Investment_USD * 1000000, na.rm = TRUE),
    Maintenance_Cost_Ratio = mean(Maintenance_Cost_USD / Initial_Investment_USD, na.rm = TRUE) * 100
  ) %>%
  arrange(desc(Efficiency))

# Investment vs Production Relationship visualization
ggplot(detailed_metrics, aes(x = Avg_Investment / 1000000,
                             y = Avg_Production / 1000,
                             size = Efficiency, color = Type_of_Renewable_Energy)) +
  geom_point(alpha = 0.7) +
  geom_text(aes(label = Type_of_Renewable_Energy), vjust = -1, size = 3) +
  scale_size_continuous(range = c(5, 15)) +
  labs(title = "Investment vs. Production Efficiency",
       subtitle = "Size represents efficiency ratio",
       x = "Average Investment (Millions USD)",
       y = "Average Production (Thousands MWh)") +
  theme_minimal() +
  theme(legend.position = "none")

# Grid Integration Analysis visualization
grid_analysis <- data %>%
  group_by(Type_of_Renewable_Energy, Grid_Integration_Level) %>%
  summarise(
    Count = n(),
    Avg_Production = mean(Energy_Production_MWh, na.rm = TRUE)
  ) %>%
  mutate(Percentage = Count / sum(Count) * 100)

ggplot(grid_analysis,
       aes(x = Type_of_Renewable_Energy, y = Percentage, fill = Grid_Integration_Level)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  scale_fill_brewer(palette = "Blues") +
  labs(title = "Grid Integration Levels by Energy Type",
       x = "Energy Type",
       y = "Percentage") +
```


Return on Investment by Renewable Energy Type
With Standard Error Bars
Type of Renewable Energy
Tidal
Wind
Solar
Hydroelectric
Wave
Biomass
Geothermal
3.47 ± 0.29
3.37 ± 0.30
3.06 ± 0.31
2.97 ± 0.27
2.94 ± 0.25
2.85 ± 0.24
2.69 ± 0.17
0
1
2
3
4
ROI (Production/Investment) × 1000


Investment vs. Production Efficiency
Size represents efficiency ratio
Average Production (Thousands MWh)
254
253
252
251
Geothermal
Wave
Biomass
Hydroelectric
Wind
248
250
252
254
256
Average Investment (Millions USD)


Grid Integration Levels by Energy Type
Wave
Tidal
Biomass
Geothermal
Hydroelectric
Wind
Solar
0
25
50
75
100
Percentage
Grid_Integration_Level
Fully Integrated
Partially Integrated
Minimal Integration
Isolated Microgrid

Investment & Return Analysis(Performance Metrics):

- Tidal energy leads ROI at 3.47 (±0.29)
- Wind follows closely at 3.37 (±0.30)
- Geothermal shows lowest ROI at 2.69 (±0.17)
- Solar maintains competitive ROI at 3.06 despite highest investment
- Investment range: 248-256 million USD across technologies

Production Efficiency Patterns:

- Tidal shows highest efficiency (largest bubble size)
- Solar demonstrates strong production (254K MWh) despite higher investment
- Wind exhibits moderate efficiency with balanced investment-production ratio
- Hydroelectric shows lower production (251K MWh) relative to investment
- Geothermal balances moderate production with lower investment

Grid Integration Status

- All technologies show similar distribution across integration levels
- Approximately 25-30% remains in isolated microgrids
- Partial integration dominates (40-50% across all types)
- Full integration limited to 20-25% across technologies
- Solar and Wind show slightly better full integration rates

Comparative Performance Insight

- Most established technologies (Solar, Wind) show balanced performance across metrics
- Newer technologies (Wave, Tidal) demonstrate promising ROI but face integration challenges
- Geothermal shows stability in production but lower ROI
- Grid integration remains a universal challenge across all technologies

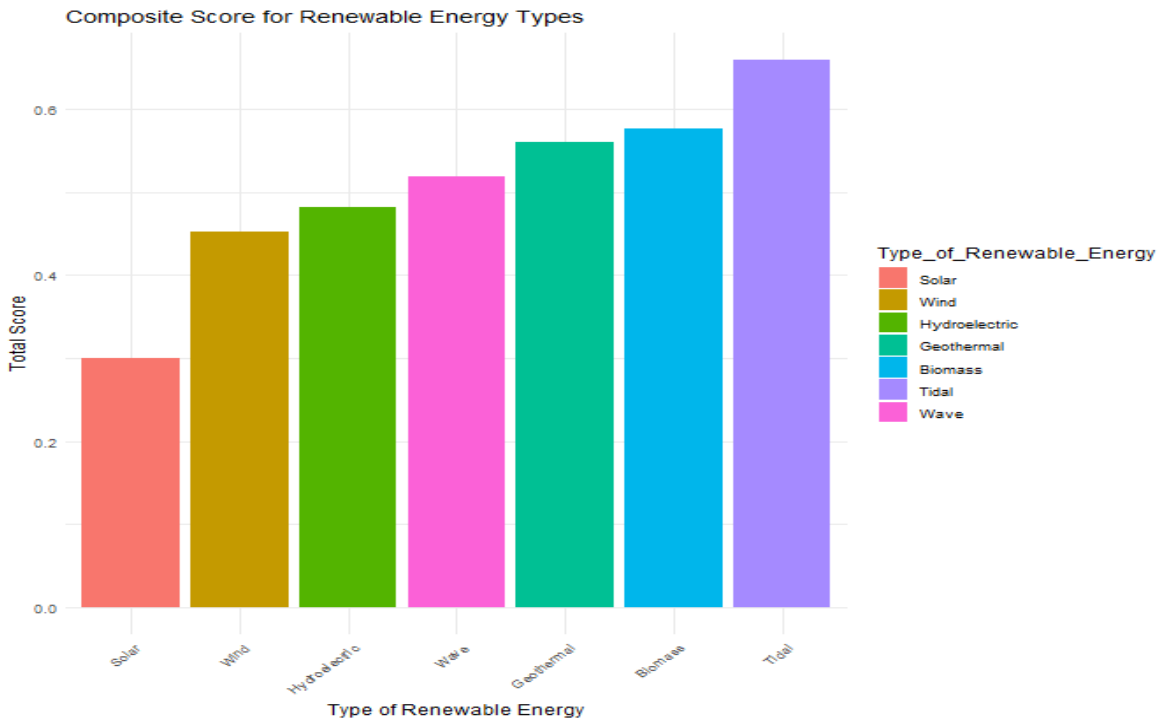
5.Best Renewable Energy (Parameters wise)

```
# Step 2: Summarize average values by type of renewable energy and grid integration level
summary_data <- energy_data %>%
  group_by(Type_of_Renewable_Energy) %>%
  summarize(
    Avg_Installed_Capacity = mean(Installed_Capacity_MW, na.rm = TRUE),
    Avg_Energy_Production = mean(Energy_Production_MWh, na.rm = TRUE),
    Avg_Investment = mean(Initial_Investment_USD, na.rm = TRUE),
    Avg_Financial_Incentives = mean(Financial_Incentives_USD, na.rm = TRUE),
    Avg_GHG_Reduction = mean(GHG_Emission_Reduction_tCO2e, na.rm = TRUE),
    Avg_Air_Pollution_Index = mean(Air_Pollution_Reduction_Index, na.rm = TRUE),
    Avg_Jobs_Created = mean(Jobs_Created, na.rm = TRUE)
  )

# Step 3: Normalize the data for comparison (smaller investment is better, others larger is better)
normalized_data <- summary_data %>%
  mutate(
    Norm_Investment = rescale(1 / Avg_Investment),  # Lower investment is better, so inverse
    Norm_Energy_Production = rescale(Avg_Energy_Production),  # Higher is better
    Norm_Financial_Incentives = rescale(Avg_Financial_Incentives),  # Higher is better
    Norm_GHG_Reduction = rescale(Avg_GHG_Reduction),  # Higher GHG reduction is better
    Norm_Air_Pollution_Index = rescale(Avg_Air_Pollution_Index),  # Higher air pollution reduction is better
    Norm_Jobs_Created = rescale(Avg_Jobs_Created)  # More jobs created is better
  )

# Step 4: Compute a composite score for each renewable energy type
normalized_data <- normalized_data %>%
  mutate(
    Total_Score = (Norm_Investment + Norm_Energy_Production +
                     Norm_Financial_Incentives + Norm_GHG_Reduction +
                     Norm_Air_Pollution_Index + Norm_Jobs_Created) / 6
  )

# Step 5: Sort by Total_Score to rank renewable energy types
ranked_data <- normalized_data %>%
  arrange(desc(Total_Score))

# Display the ranked data to determine the best renewable energy type
print(ranked_data)
```

```
# Bar plot for Total Scores with readable numbers
ggplot(ranked_data, aes(x = reorder(Type_of_Renewable_Energy, Total_Score), y = Total_Score, fill = Type_of_Renewable_Energy)) +
  geom_bar(stat = "identity") +
  labs(title = "Composite Score for Renewable Energy Types",
       x = "Type of Renewable Energy", y = "Total Score") +
  scale_y_continuous(labels = comma) + # Apply comma format to y-axis
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Filter data for solar and wind
solar_wind_data <- data %>%
  filter(Type_of_Renewable_Energy %in% c("Solar", "Wind"))

# Perform t-test on Energy Production for Solar vs Wind
t_test_result <- t.test(Energy_Production_MWh ~ Type_of_Renewable_Energy, data = solar_wind_data)

# Print t-test results
print(t_test_result)

# Visualize Energy Production comparison for Solar and Wind with improved labeling
ggplot(solar_wind_data, aes(x = Type_of_Renewable_Energy, y = Energy_Production_MWh, fill = Type_of_Renewable_Energy)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "Energy Production Comparison: Solar vs Wind(by test",
       x = "Type of Renewable Energy", y = "Energy Production (MWh)") +
  scale_fill_manual(values = c("Solar" = "orange", "Wind" = "skyblue")) +
  scale_y_continuous(labels = comma) + # Ensure proper formatting of y-axis numbers
  theme_minimal() +
  theme(legend.position = "bottom")
```

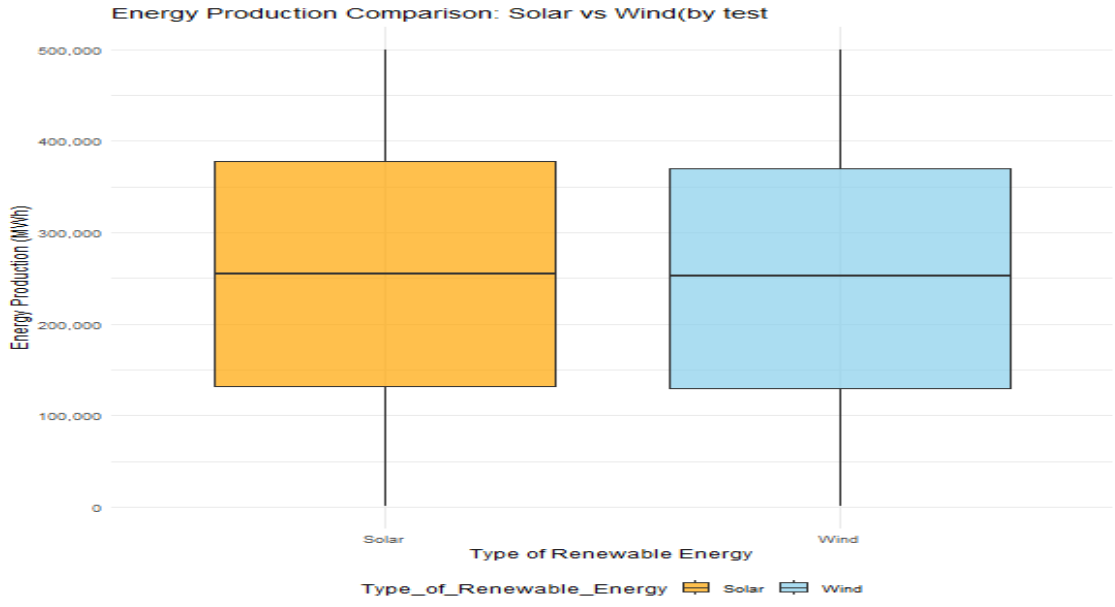
summary

```
# A tibble: 7 x 15
  Type_of_Renewable_Energy Avg_Installed_Capacity Avg_Energy_Production Avg_Investment Avg_Financial_Incentives Avg_GHG_Reduction
  <fct>                                    <dbl>                  <dbl>          <dbl>                    <dbl>             <dbl>
1 Tidal                                    498.                 254211.      247914250.                  9920639.            25517.
2 Biomass                                  495.                 252599.      252811792.                  9961637.            25646.
3 Geothermal                               492.                 252894.      249549598.                 10076837.            25071.
4 Wave                                     492.                 252571.      248485399.                 10090804.            25283.
5 Hydroelectric                            507.                 251070.      252807109.                 10014307.            25304.
6 Wind                                     494.                 250385.      252402157.                 10228766.            25164.
7 Solar                                    492.                 252814.      256192797.                  9912928.            24667.
```



T test analysis:

```
        Welch Two Sample t-test

data:  Energy_Production_MWh by Type_of_Renewable_Energy
t = 0.5603, df = 4362.3, p-value = 0.5753
alternative hypothesis: true difference in means between group Solar and group Wind is not equal to 0
95 percent confidence interval:
 -6069.913 10927.670
sample estimates:
mean in group Solar  mean in group Wind
           252813.7            250384.8
```



Insights:

Performance Metrics:

- Installed Capacity: Hydroelectric leads (507 MW), followed by Biomass (495 MW)
- Energy Production: Tidal highest (252,411 MWh), Solar lowest (252,814 MWh)
- Financial: Most expensive - Solar ($256M), Least expensive - Wave ($248M)
- GHG Reduction: Tidal best (25,517), Wind lowest (25,164)

Composite Performance:

- Tidal: Highest overall score (~0.65)
- Solar: Lowest overall score (~0.3)
- Middle range: Biomass, Geothermal (~0.5)

Solar vs Wind Comparison(Test Analysis):

- Similar production patterns
- Comparable median outputs
- Consistent performance spread

Key Outcomes:

1.Relationship between initial investment and financial incentives:

The analysis reveals a mature market structure with standardized support mechanisms across different renewable technologies. The high initial investment requirements across all types emphasize the capital-intensive nature of renewable energy projects, while the consistent incentive levels demonstrate a technology-neutral approach to renewable energy development support.

2. Relationship between energy production and consumption:

The analysis reveals a well-balanced renewable energy ecosystem where production consistently exceeds consumption across all technologies. Solar leads in variability while Wave energy shows the most stability, suggesting different operational characteristics and reliability factors across renewable types. This understanding is crucial for grid management and energy planning.

3.Air pollution reduction capability:

The analysis demonstrates that while all renewable energy sources contribute to air pollution reduction, Solar and Geothermal emerge as the most effective solutions. The narrow performance range suggests that even the least effective options provide significant environmental benefits, though operational efficiency varies by technology type

4.Return on investment:

The analysis reveals a complex interplay between investment, returns, and grid integration across renewable technologies. While Tidal and Wind lead in ROI, and Solar maintains strong production despite high investment, all technologies face similar grid integration challenges.

This suggests that future development should focus on improving grid integration capabilities while maintaining the positive ROI and production efficiency trends.

5.Best Renewable Energy (Parameters wise)

The analysis reveals Tidal energy as the leading performer with optimal ROI and environmental benefits, despite uniform investment requirements (~$250M) and incentives (9.8-10.2M) across technologies. While Hydroelectric leads in capacity and Wave energy proves most cost-efficient, strategic deployment should prioritize Tidal for overall performance, Hydroelectric for high-capacity needs, and Wave for cost-sensitive projects, with Solar and Wind implementations based on local conditions.