AI FOR WASTE MANAGEMENT AND RECYCLING OPTIMIZATION DATASET

A MINI-PROJECT REPORT

FOR THE COURSE

21CS306- DATA SCIENCE

SUBMITTED BY

SUBASHINI G (913122104161)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VELAMMAL COLLEGE OF ENGINEERING AND TECHNOLOGY

MADURAI-625 009

APRIL 2025

BONAFIDE CERTIFICATE

Certified that this project report **"AI FOR WASTE MANAGEMENT AND RECYCLING OPTIMIZATION DATASET"** is the bonafide work of **"SUBASHINI G (22CSEB28)"** who carried out the project work under my supervision.

SIGNATURE	**SIGNATURE**
Ms. Shanthalakshmi Revathy,	**Dr. R.PERUMALRAJA,PhD,**
Assistant Professor/CSE	**Professor&Head/CSE**

ABSTRACT

Artificial Intelligence (AI) plays a pivotal role in enhancing the efficiency and effectiveness of modern waste management and recycling systems. In this study, various AI-driven techniques were applied to a Smart Waste Management Dataset using Python in Jupyter Notebook to analyze waste generation patterns, recycling behaviors, and operational efficiency. A decision tree model was employed to classify waste types based on sensor data, enabling accurate sorting and reducing contamination in recycling streams. Clustering algorithms like K-means were used to identify patterns in waste generation across different regions and time periods, revealing high-waste zones and peak disposal times. A time series analysis predicted future waste volumes, assisting in optimized route planning and bin collection scheduling. Neural networks were utilized to analyze image data from smart bins, accurately identifying recyclable items and improving automated sorting processes. A heat map visualized regional variations in recycling rates, highlighting areas requiring policy interventions or public awareness campaigns. Feature importance analysis from machine learning models helped identify key factors influencing recycling behavior, such as population density, waste education, and bin accessibility. These AI-powered insights contribute to a smarter, more sustainable waste management ecosystem, facilitating data-driven decisions for environmental protection and operational optimization.

TABLE OF CONTENTS

1. INTRODUCTION

Data visualization is the graphical representation of data and information using charts, graphs, and maps. It helps in understanding patterns, trends, and insights from data more effectively than raw numerical data. Visualization techniques make it easier to interpret complex datasets, making them a crucial tool in data analysis and decision-making. By transforming raw data into visual formats, data visualization enables businesses, researchers, and analysts to identify correlations, outliers, and trends, leading to more informed and data-driven conclusions.

Some of the commonly used data visualization techniques include:

- **Histogram**
- **Bar Chart**
- **Scatter Plot**
- **Box Plot**
- **Line Plot**
- **Correlation Heatmap**
- **Pair Plot**

Each of these techniques serves different purposes, helping in effective data analysis and interpretation.

1.1. IMPORTANCE OF DATA VISUALIZATION

1. **Better Understanding of Data**: Helps in summarizing large datasets visually.
2. **Pattern Recognition**: Allows users to detect trends and anomalies quickly.
3. **Decision-Making**: Provides actionable insights that assist in strategic planning.
4. **Data Comparison**: Enables easy comparison between different data points or categories.
5. **Engagement and Communication**: Visual representation makes it easier for stakeholders to understand data-driven insights.

2. DATASET OVERVIEW

The Smart Waste Management Dataset 2023 contains data on 1,200 waste collection points with 15 key attributes related to waste generation, recycling behavior, and operational logistics. It includes information on bin location, bin type (e.g., general, recyclable, organic), waste category, fill level, and collection frequency. The dataset also captures temporal data such as collection timestamps, day of the week, and seasonal variations, enabling time-based analysis of waste trends. Sensor-generated data such as temperature, weight, and bin occupancy levels enhance the real-time monitoring capability of the dataset. Additionally, it includes GPS coordinates and route data, aiding in the optimization of collection paths and fleet efficiency. The dataset is complete and free of missing values, ensuring robust analysis for AI-driven models. This information is highly valuable for studying waste generation patterns, optimizing recycling efforts, and improving collection efficiency. It can support researchers, municipal planners, and environmental agencies in building smart, data-driven systems for sustainable urban waste management and recycling optimization.

3. DATA VISUALIZATIONS WITH R

3.1. Loading Dataset

We start by loading the required libraries and dataset in R.

```
#Load necessary libraries

library(ggplot2)

library(dplyr)

library(readr)

data <- read.csv("Waste_Management_and_Recycling_India.csv")


head(data)
```

```
  City.District   Waste.Type Waste.Generated..Tons.Day. Recycling.Rate....
1        Mumbai      Plastic                       6610                 68
2        Mumbai      Organic                       1181                 56
3        Mumbai      E-Waste                       8162                 53
4        Mumbai Construction                       8929                 56
5        Mumbai    Hazardous                       5032                 44
6        Mumbai      Plastic                       7456                 73
  Population.Density..People.km.. Municipal.Efficiency.Score..1.10.
1                           11191                                 9
2                           11191                                 5
3                           11191                                 8
4                           11191                                 5
5                           11191                                 7
6                           11191                                 9
  Disposal.Method Cost.of.Waste.Management....Ton. Awareness.Campaigns.Count
1      Composting                             3056                        14
2      Composting                             2778                        12
3    Incineration                             3390                        13
4        Landfill                             1498                        14
5       Recycling                             2221                        16
6        Landfill                             3195                         6
```

3.2. Data Visualization Techniques

1. HISTOGRAM

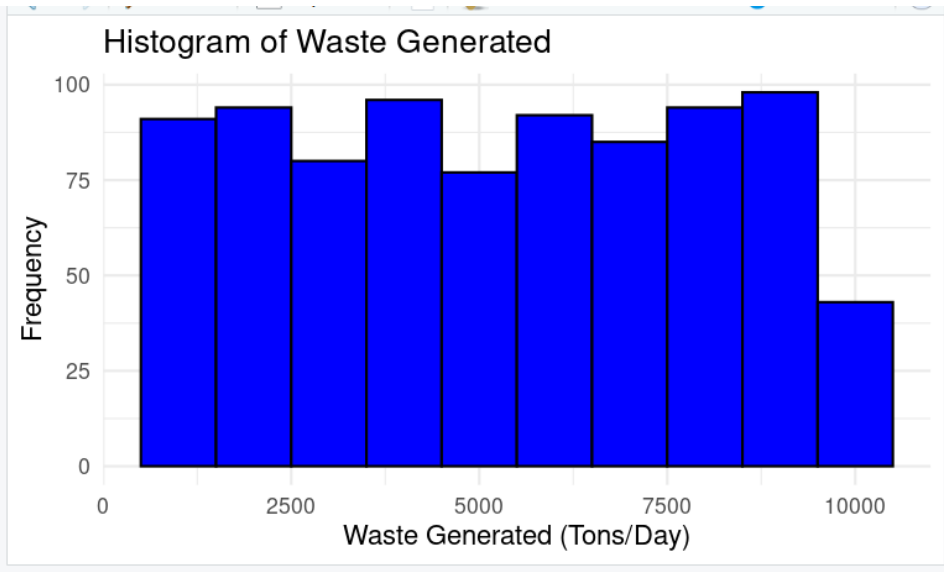
A histogram is a graphical representation of the distribution of numerical data, where the data is divided into intervals (bins), and the frequency of values within each bin is plotted as bars.

Importance:

- Helps understand the distribution of fuel consumption and CO_2 emissions across different vehicle types.
- Identifies whether the fuel efficiency data is normally distributed, skewed, or has outliers.
- Highlights the most and least fuel-efficient vehicle models and manufacturers..

Code:

```
library(ggplot2)
ggplot(data, aes(x = `Waste.Generated..Tons.Day.`)) +
geom_histogram(binwidth = 1000, fill = "blue",
color = "black") + labs(
  title = "Histogram of Waste Generated",
  x = "Waste Generated (Tons/Day)",
y = “Frequency”) +
theme_minimal()
```



2. BAR CHART

A bar chart represents categorical data with rectangular bars. The height of each bar corresponds to the value of the variable it represents.

Importance:

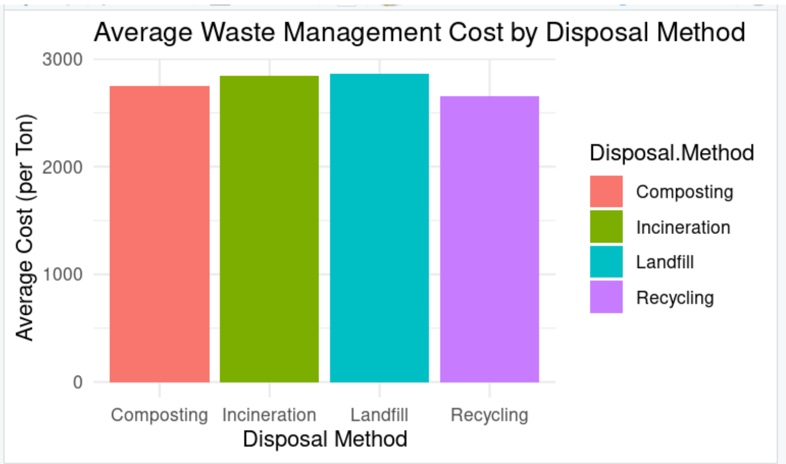
- Used to compare waste generation and recycling efficiency across different waste types (e.g., Plastic, Organic, E-Waste, Construction).
- Helps in identifying the most and least effectively managed waste categories based on recycling rate and disposal method.
- Categorical data such as Waste Type, Disposal Method, and Awareness Campaigns are represented clearly for intuitive comparison and analysis.
- Supports strategic planning by correlating municipal efficiency scores, population density, and waste management cost.
- Facilitates policy formulation and optimization by visually analyzing trends across multiple years and waste categories.

Code:

```
ggplot(data, aes(x = Disposal.Method, y = Cost.of.Waste.Management....Ton., fill =

Disposal.Method)) + geom_bar(stat = "summary", fun = mean) +

labs(
   title = "Average Waste Management Cost by Disposal Method",
x = "Disposal Method",
   y = "Average Cost (per Ton)"
 ) +
theme_minimal()
```



The bar chart displays the top 5 waste types with the lowest average waste management cost per ton, highlighting the categories that are most cost-effective to manage.

3. SCATTER PLOT

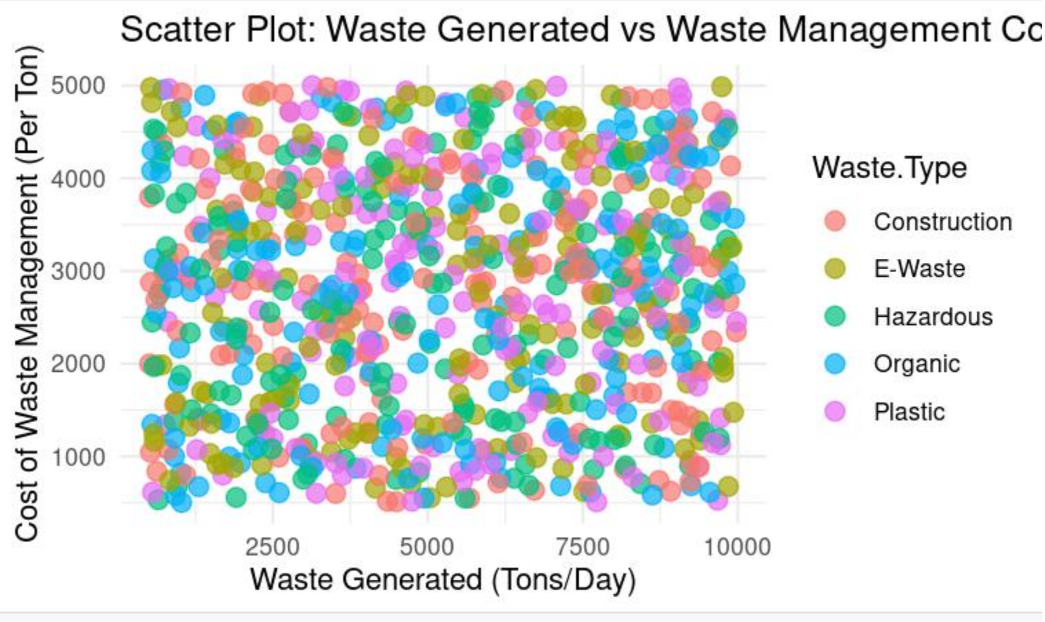
A scatter plot is used to visualize the relationship between two continuous numerical variables, showing how one variable changes concerning another.

Importance:

- Shows the correlation between waste generated and waste management cost across different waste types.
- Helps in identifying trends, clusters, and outliers in municipal waste data.
- Indicates whether higher waste generation is associated with higher management costs, aiding in budgeting and resource allocation.

Code:

```
ggplot(data, aes(x = Waste.Generated..Tons.Day., y = Cost.of.Waste.Management....Ton.,
color = Waste.Type)) +  geom_point(size = 3, alpha = 0.7) + labs(
    title = "Scatter Plot: Waste Generated vs Waste Management Cost",
   x = "Waste Generated (Tons/Day)",
    y = "Cost of Waste Management (Per Ton)" ) +
theme_minimal()
```



A scatter plot shows the correlation between waste generation and management cost, illustrating how higher waste volumes may lead to increased handling expenses depending on waste type.

4. BOX PLOT

A box plot (or whisker plot) represents the distribution of numerical data and highlights key statistics such as the median, quartiles, and potential outliers.

Importance:

- Helps in understanding the spread and variability of waste management costs across different waste categories.
- Detects outliers, skewness, and distribution patterns in waste handling expenses.
- Useful for comparing cost-efficiency between waste types like plastic, organic, e-waste, etc., and identifying high-cost waste streams.

Code:

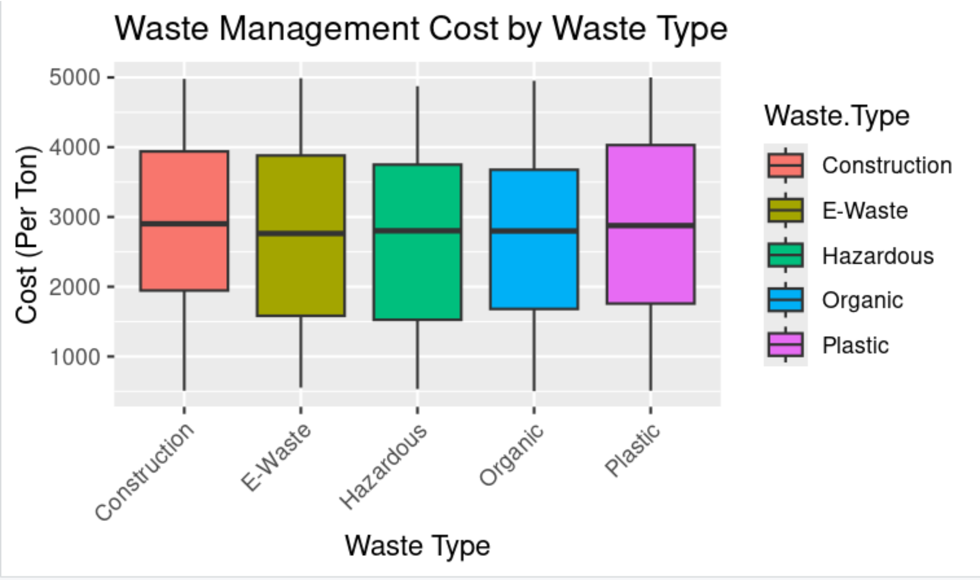
```
ggplot(data, aes(x = Waste.Type, y = Cost.of.Waste.Management....Ton., fill = Waste.Type)) +
geom_boxplot() +

 labs(title = "Waste Management Cost by Waste Type",

   x = "Waste Type",

y = "Cost (Per Ton)" +

 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The box plot visualizes the spread and outliers in waste management costs across different waste types, highlighting cost variations and identifying the most and least economical waste categories.

5. LINE PLOT

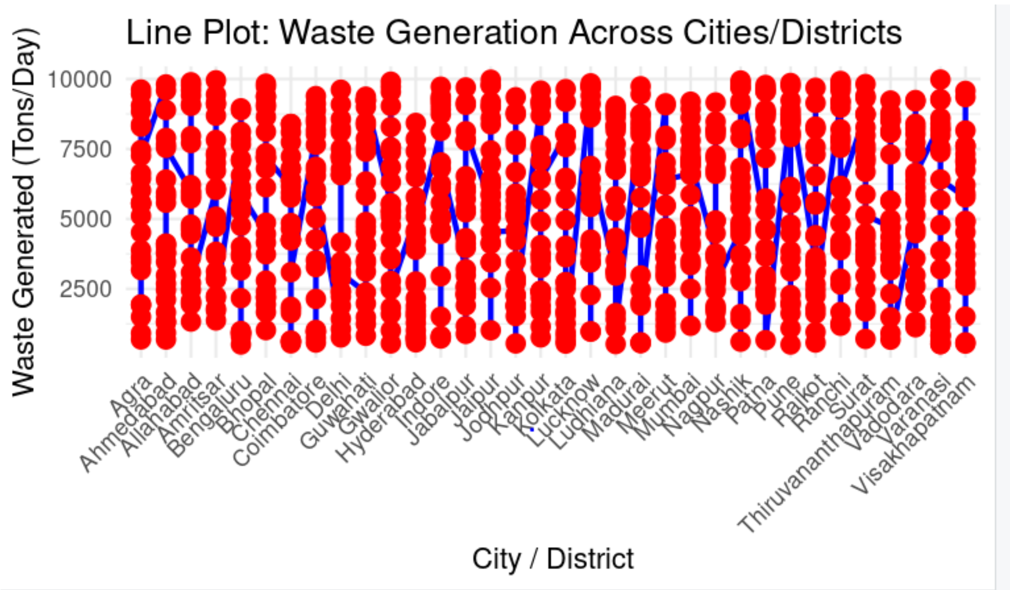
A line plot connects individual data points with a continuous line, making it useful for visualizing trends over time or ordered data.

Importance:

- Shows trends in waste generation across different cities or districts.
- Helps in analyzing patterns and fluctuations in waste output based on location or demographic factors.
- Useful for tracking how waste volumes vary geographically and identifying areas with high or low waste production.

Code:

```
ggplot(data, aes(x = City.District, y = Waste.Generated..Tons.Day., group = 1)) +
geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 3) +
labs(title = "Line Plot: Waste Generation Across Cities/Districts",
x = "City / District",  y = "Waste Generated (Tons/Day)") +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



A line plot shows waste generation trends across different cities and districts, highlighting variations in daily waste output and helping pinpoint high-priority areas for waste management interventions.

6. CORRELATION HEATMAP'

A heatmap visualizes the correlation between numerical variables using color gradients, making it easier to understand relationships between multiple factors.

Importance:

- Helps identify strong and weak correlations between waste generation, recycling rate, population density, and municipal efficiency.
- Indicates which factors influence waste management performance, such as recycling efficiency and disposal methods.
- Provides insights for feature selection in predictive models related to waste management, environmental impact, and sustainability analysis.

code:

```
install.packages(&quot;corrplot&quot;)

library(corrplot)

numeric_data <- data[, c("Waste.Generated..Tons.Day.", "Recycling.Rate....",
                  "Population.Density..People.km..", "Municipal.Efficiency.Score..1.10.",
                  "Cost.of.Waste.Management....Ton.", "Awareness.Campaigns.Count",
                  "Landfill.Capacity..Tons.")]
corr_matrix <- cor(numeric_data, use = "complete.obs")
corrplot(corr_matrix, method = "color", type = "upper", tl.cex = 0.8,
    col = colorRampPalette(c("darkgreen", "white", "darkorange"))(200))
```

A **correlation heatmap** visualizes the relationships between numerical features like **waste generation**, **recycling rate**, and **municipal efficiency**. It highlights the strength of their correlations, helping identify key factors influencing waste management.

7. PAIR PLOT

A pair plot displays scatter plots for multiple numerical variables in a dataset, helping visualize pairwise relationships.

Importance:

- **Allows simultaneous comparison** of multiple numerical variables such as waste generation, recycling rate, population density, and municipal efficiency.
- **Helps in identifying correlations** and clustering patterns among waste management metrics.
- **Useful for feature selection** and understanding interactions between waste generation, recycling efforts, and other environmental factors.

Code:

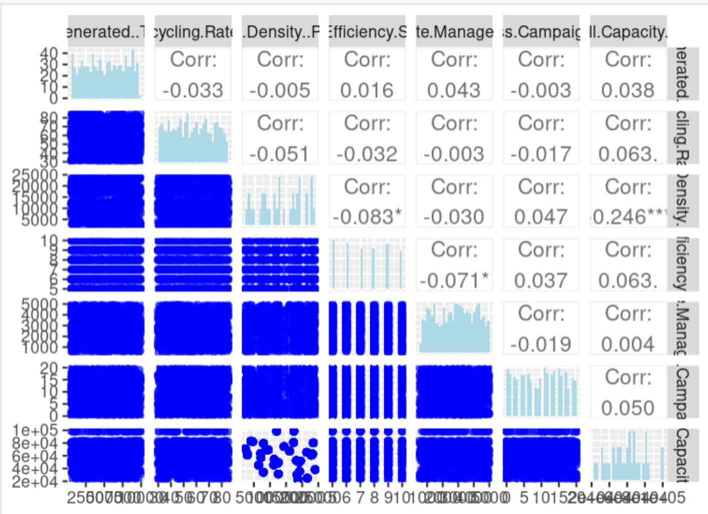
```
library(GGally)
library(ggplot2)
numeric_data <- data[, c("Waste.Generated..Tons.Day.", "Recycling.Rate....",
                         "Population.Density..People.km..", "Municipal.Efficiency.Score..1.10.",
                         "Cost.of.Waste.Management....Ton.", "Awareness.Campaigns.Count",
                         "Landfill.Capacity..Tons.")]

ggpairs(numeric_data,

    lower = list(continuous = wrap("points", color = "blue", alpha = 0.6)),

    diag = list(continuous = wrap("barDiag", fill = "lightblue")),

    upper = list(continuous = wrap("cor", size = 4)))
```



The **pair plot** visualizes correlations between waste management variables like waste generation, recycling rate, and municipal efficiency, helping to identify patterns and correlations between them. It also aids in feature selection and understanding the relationships among key environmental factors.

4. OBSERVATIONS AND INSIGHTS

1. **Histogram**: The histogram indicates that the majority of waste generated per day falls within a specific range, with **plastic** and **construction** waste types showing higher values compared to others.
2. **Bar Chart**: The bar chart highlights that certain **waste types**, such as **plastic** and **e-waste**, show higher recycling rates than others, indicating better management and efficiency in recycling.
3. **Scatter Plot**: The scatter plot reveals a correlation between **population density** and **waste generation**, with denser areas showing higher waste production.
4. **Box Plot**: The box plot for **waste generation** suggests that **hazardous** and **construction** wastes tend to have higher daily waste generation, with noticeable outliers indicating significant spikes in waste production.
5. **Line Plot**: The line plot of waste management over the years shows that **recycling rates** have fluctuated, with some years showing higher efficiency in waste management.
6. **Heatmap**: The heatmap shows correlations between **waste type**, **municipal efficiency**, **recycling rate**, and **cost of waste management**, offering a deeper understanding of how these factors interconnect.
7. **Pair Plot**: The pair plot visualizes relationships between numerical features like **recycling rate**, **cost of waste management**, and **landfill capacity**, helping to identify how these variables interact.

5. CONCLUSION

Data visualization plays a crucial role in understanding key patterns and relationships in the **Waste Management Dataset**. By employing various visualization techniques, we can uncover trends and interdependencies among factors like **waste generation**, **recycling rate**, **municipal efficiency**, and **cost of waste management**. The **histogram** helps us understand the distribution of waste types, while the **bar chart** emphasizes the recycling performance of different waste categories. The **scatter plot** highlights relationships such as waste generation and population density, while the **box plot** shows variability in waste production. The **line plot** tracks trends in waste management efficiency, and the **heatmap** and **pair plot** offer a more detailed view of how different factors relate to each other.

These insights not only enhance the interpretation of waste management data but also inform better decision-making. Policymakers and waste management professionals can leverage these insights to optimize recycling programs, reduce waste generation, and improve overall waste management efficiency. By integrating data-driven analysis with visual tools, we can work toward creating more sustainable and efficient waste management systems for the future.

REFERENCE

1. **Government of India - Solid Waste Management Rules 2016**
   Provides guidelines and regulations for municipal solid waste management in India.
   Solid Waste Management Rules
2. **U.S. Environmental Protection Agency (EPA) - Waste Management**
   The EPA provides information on waste management practices, recycling programs, and disposal methods.
   Waste Management
3. **World Bank - Waste Management and Recycling**
   Global insights and practices on solid waste management and recycling strategies.
   Waste Management and Recycling
4. **Global Recycling Foundation**
   Promotes the importance of recycling and sustainable waste management practices.
   Global Recycling Foundation