Enhancing Resource Management

Assessing Local Recycling Efficiency and Waste Management Impact in NYC Community Districts

George Mason University

AIT-580-007 | Prof. Dr. Harry Foxwell

Department of Data Analytics and Engineering

George Mason University, Virginia

Mano Harsha Sappa

G01459796

Msappa@gmu.edu

Abstract: This project conducts a comprehensive analysis of New York City's recycling landscape, specifically focusing on the recycling diversion and capture rates within its Community Districts. By examining data from multiple fiscal years, the study evaluates various metrics, including recycling efficiency, paper capture efficiency, metal/glass/plastic (MGP) capture efficiency, and total capture efficiency for each district. The analysis not only assesses the effectiveness of the Department of Sanitation's (DSNY) recycling programs but also explores the correlation between these rates and overall recycling performance. Through an in-depth investigation into the complexities of waste management, the study offers valuable insights into optimizing recycling initiatives and reducing waste sent to landfills. Furthermore, it highlights the importance of ongoing monitoring and evaluation in shaping sustainable waste management practices, thus contributing to a greener and more resilient urban environment.

Keywords: Recycling efficiency, Municipal solid waste, Capture rate, Waste management, Department of Sanitation (DSNY), New York City, Community Districts, Sustainability, Waste diversion, Recycling programs, Environmental impact, Urban sustainability, Resource optimization, Landfill reduction, Sustainable cities.

I. Introduction

Since waste management is a critical aspect of urban sustainability, understanding the effectiveness of recycling efforts is paramount. In this project, we delve into the recycling landscape of New York City, focusing on the Recycling Diversion and Capture Rates within its Community Districts. The Recycling Diversion rate indicates the percentage of total municipal solid waste that is diverted from landfills through recycling, while the Capture Rate measures the efficiency of recycling specific materials such as paper, metal, glass, and plastic.

Through the analysis of data spanning multiple fiscal years, we aim to evaluate the performance of the Department of Sanitation's (DSNY) recycling programs and identify areas for improvement. By examining metrics such as Recycling Efficiency, Paper Capture Efficiency, Metal/Glass/Plastic (MGP) Capture Efficiency, and Total Capture Efficiency for each Community District, we seek to gain insights into the effectiveness of recycling initiatives across different neighborhoods.

It is worth noting that since 2013, DSNY no longer utilizes capture rate information, which adds a contextual layer to our analysis. Nevertheless, by leveraging available data and considering the evolving landscape of waste management practices, this project endeavors to provide actionable recommendations for enhancing recycling efforts and promoting sustainable waste management practices in New York City.

His study delves into the trends spanning from 2016 to 2019, focusing on recycling diversion rates and capture rates across NYC's community districts. By analyzing datasets provided by NYC Open Data, encompassing metrics such as Recycling Efficiency, Paper Capture Efficiency, and Metal/Glass/Plastic (MGP) Capture Efficiency, insights into the performance of recycling programs can be gleaned. Additionally, exploration of seasonal variations, as indicated by Fiscal Month Number and Month Name, provides a nuanced understanding of temporal patterns influencing recycling outcomes.

II. Research Questions

1. How does the mean recycling efficiency in New York City vary by month over the fiscal years 2016 to 2019?

2. What are the potential environmental and economic implications of the observed trends in recycling diversion and capture rates in New York City?

3. How do fluctuations in seasonality, as denoted by Fiscal Month Number and Month Name, affect the Diversion Rate-Total and Capture Rates, shedding light on the temporal patterns of recycling performance?

III.Dataset Analysis Questions

- What is the average recycling efficiency for each district in New York City?

- Which month had the highest total capture efficiency, and what was the value?

- How does recycling efficiency vary across different fiscal years?

- Which zone has the highest average total capture efficiency?

- Can you identify any correlation between paper capture efficiency and MGP (Metal, Glass, Plastic) capture efficiency?

- What is the overall recycling performance trend over the years?

- Are there any districts where the total capture efficiency exceeds a certain threshold (e.g., 50%) consistently?

- How does recycling efficiency change month by month?

- Is there a significant difference in recycling efficiency between different zones?

- Can we identify any seasonal trends in recycling diversion and capture rates?

The implications of observed trends in recycling diversion and capture rates extend beyond environmental considerations, encompassing potential economic impacts and policy implications. This study aims to elucidate the multifaceted dimensions of NYC's recycling landscape, informing stakeholders and policymakers for informed decision-making and sustainable waste management practices.

Keywords: Waste management, Urban sustainability, Recycling, New York City, Recycling Diversion Rate, Capture Rate, Community Districts, Department of Sanitation (DSNY), Recycling Efficiency, Paper Capture Efficiency, Metal/Glass/Plastic (MGP) Capture Efficiency, Total Capture Efficiency, NYC Open Data, Seasonal variations, Fiscal Month Number, Month Name, Environmental implications, Economic implications, Policy implications, Sustainable waste management practices.

IV. Dataset Analysis Questions

- What is the average recycling efficiency for each district in New York City?

- Which month had the highest total capture efficiency, and what was the value?

- How does recycling efficiency vary across different fiscal years?

- Which zone has the highest average total capture efficiency?

- Can you identify any correlation between paper capture efficiency and MGP (Metal, Glass, Plastic) capture efficiency?

- What is the overall recycling performance trend over the years?

- Are there any districts where the total capture efficiency exceeds a certain threshold (e.g., 50%) consistently?

- How does recycling efficiency change month by month?

- Is there a significant difference in recycling efficiency between different zones?

- Can we identify any seasonal trends in recycling diversion and capture rates?

The implications of observed trends in recycling diversion and capture rates extend beyond environmental considerations, encompassing potential economic impacts and policy implications. This study aims to elucidate the multifaceted dimensions of NYC's recycling landscape, informing stakeholders and policymakers for informed decision-making and sustainable waste management practices.

Keywords: *Waste management, Urban sustainability, Recycling, New York City, Recycling Diversion Rate, Capture Rate, Community Districts, Department of Sanitation (DSNY), Recycling Efficiency, Paper Capture Efficiency, Metal/Glass/Plastic (MGP) Capture Efficiency, Total Capture Efficiency, NYC Open Data, Seasonal variations, Fiscal Month Number, Month Name, Environmental implications, Economic implications*

V. Problem Statement

In New York City, managing waste effectively and encouraging recycling are crucial for keeping the city clean and sustainable. However, understanding how well recycling efforts are working requires a deeper look into the rates of recycling diversion and capture across different areas of the city.

. I want to study how recycling efficiency changes in New York City each month from 2016 to 2019. By looking at recycling data month by month, I hope to see if there are any patterns or trends over those years. The main goal is to find out if recycling rates go up and down throughout the year.

. Environmental and Economic Impact: Understanding the effects of recycling on the environment and the economy is important. By analyzing trends in recycling rates, we can see how they might be affecting things like pollution levels, natural resources, and the city's finances.

. Seasonal Patterns: Recycling rates might change throughout the year due to factors like weather, holidays, or special events. By studying these fluctuations, we can uncover seasonal patterns in recycling behavior and better plan for managing waste during different time of the year.

VI. Ethical Considerations in Recycling Data Analysis

1. Privacy Protection and Transparency:

Respecting privacy rights is paramount in our data analysis process. We are committed to safeguarding the confidentiality of individuals and organizations mentioned in the recycling dataset. Any personal information will be handled with utmost care and will not be shared without consent. Transparency is key to building trust with stakeholders, and we will clearly document our data collection, analysis, and reporting methods. By being transparent about our processes, we uphold accountability and ensure that stakeholders understand how their data is being used [1]."

2. Avoiding Harm and Ensuring Honesty:

Our analysis prioritizes preventing harm and promoting data security. We are mindful of the potential risks associated with misusing or misinterpreting the data and take proactive measures to mitigate them. By adhering to ethical principles and maintaining honesty in our analysis, we aim to provide accurate and reliable insights into recycling practices. Transparency and integrity are at the core of our approach, ensuring that our findings are trustworthy and beneficial to all stakeholders involved [2]."

3. Regulatory Compliance and Legal Considerations:

In addition to ethical considerations, our analysis adheres to relevant laws and regulations governing data privacy, intellectual property rights, and environmental protection. We ensure compliance with data protection laws by obtaining consent for data collection and processing personal data lawfully and securely. Respecting intellectual property rights, particularly those of the Department of Sanitation of New York City (DSNY), is essential, and proper attribution will be given. Furthermore, we comply with environmental regulations and waste management policies to ensure legal compliance in our analysis [2]."

VII. Social Responsibility and Impact Assessment:

As part of our ethical framework, we consider the social responsibility of our analysis and its potential impact on society and the environment. We aim to contribute positively to societal well-being and environmental sustainability by promoting efficient waste management and resource conservation. Our analysis prioritizes fairness, equity, and non-discrimination, considering the diverse needs and perspectives of all stakeholders involved. We conduct impact assessments to evaluate the broader implications of our findings and ensure that our analysis aligns with ethical principles and societal values.

VIII. Data Type Summary

Nominal: Nominal data consists of categories with no inherent order or ranking. It's used for labeling variables without any quantitative value. Example: Zone names (e.g., "Brooklyn North", "Manhattan") [3]."

Ordinal: Ordinal data has a natural order but the intervals between categories may not be uniform. It represents categories with a relative ranking or order. Example: Month names (e.g., "January", "February").

Interval: Interval data represents values where differences are meaningful but there's no true zero point. It has a fixed unit of measurement with equal intervals between points. Example: Fiscal Year (e.g., 2016, 2017, 2019).

Ratio: Ratio data has a true zero point and meaningful ratios between values. It represents quantities where ratios are meaningful, and operations like multiplication and division are valid. Example: Recycling Efficiency, Capture Efficiency (e.g., percentages).

IX. Dataset

Final_Project_Ait_580.cleaned_recycling_diversion_capture_rates

Description: The dataset contains information on recycling diversion and capture rates across different zones and districts in New York City. It provides insights into the efficiency of recycling operations, focusing on various materials such as paper, metal, glass, and plastic. The data spans from the fiscal years 2016 to 2019, with records for each month within this timeframe.

1. Zone: Represents the geographic zone where the district is located (e.g., Brooklyn North, Brooklyn South, Bronx, Manhattan, Queens East, Queens West, Staten Island).

2. District: Specifies the specific district within the zone (e.g., BKN01, BKN02, BX01, MN01, QE07, QW01, SI01).

3. Fiscal Year: Indicates the fiscal year during which the data was recorded. The values range from 2016 to 2019.

4. Month Name: Specifies the month during which the data was recorded.

5. Recycling Efficiency: Indicates the overall efficiency of recycling operations.

6. Paper Capture Efficiency: Measures the efficiency of capturing and recycling paper materials.

7. MGP Capture Efficiency: Measures the efficiency of capturing and recycling Metal, Glass, and Plastic materials.

8. Total Capture Efficiency: Represents the overall efficiency of capturing and recycling all materials.

9. Difference: Represents any calculated differences or variances.

10. Overall Recycling Performance: Provides an overall performance score for recycling operations.

11. Fiscal Date: Specifies the fiscal date during which the data was recorded.

X. Comprehensive Analysis of Recycling Data

This data analysis project embarked on a thorough examination of recycling diversion and capture rates, employing a blend of Python, R, SQL, and AWS EC2/S3 technologies. Initially, Python was leveraged to preprocess the dataset, where meticulous steps were taken to handle missing values and simplify column names. Cumbersome column titles were transformed into more intuitive descriptors, enhancing clarity and understanding. Moreover, redundant columns were eliminated to streamline the dataset for subsequent analysis.

Following the preprocessing phase, the cleaned dataset was stored and utilized for further investigation. In R, the dataset underwent comprehensive scrutiny to address specific research questions and conduct exploratory analysis. Visualizations were crafted to illustrate trends and patterns effectively. Additionally, a series of supplementary questions were explored to provide a holistic perspective on the dataset's nuances and characteristics.

Transitioning to SQL, targeted queries were executed to delve deeper into specific aspects of the dataset, unveiling insightful findings. AWS S3 buckets were employed to facilitate data storage and sharing of visualizations, fostering collaboration and accessibility. Through this multifaceted approach, the project exemplified a structured and systematic methodology for unraveling insights from recycling data, culminating in actionable conclusions to inform waste management strategies and sustainability initiatives.

XI. Data Preprocessing and Transformation Using Python

1. Data Cleaning

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        %matplotlib inline
```

2. Loading and Initial Exploration of Recycling Diversion and Capture Rate Dataset

```
In [2]: data = pd.read_csv('/Users/manoharshasappa/Desktop/AIT 580/Final Project Ait 580/Recycling_Diversion_Capture_RateS.
        data.head()
```

Out[2]:

	Zone	District	Fiscal Month Number	Fiscal Year	Month Name	Diversion Rate-Total (Total Recycling / Total Waste)	Capture Rate-Paper (Total Paper / Max Paper)	Capture Rate-MGP (Total MGP / Max MGP)	Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100
0	Brooklyn North	BKN01	10	2019	April	14.7	44.9	43.0	44.1
1	Brooklyn North	BKN02	10	2019	April	20.0	34.2	57.9	41.2
2	Brooklyn North	BKN03	10	2019	April	12.2	33.5	44.9	38.2
3	Brooklyn North	BKN04	10	2019	April	15.5	35.2	68.5	48.8
4	Brooklyn North	BKN05	10	2019	April	10.1	22.3	45.1	31.5

3. Checking Data Types and Information of Recycling Diversion and Capture Rate Dataset

```
In [4]: data.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 2832 entries, 0 to 2831
        Data columns (total 9 columns):
         #   Column                                                                         Non-Null Count  Dtype
        ---  ------                                                                         --------------  -----
         0   Zone                                                                           2832 non-null   objec
        t
         1   District                                                                       2832 non-null   objec
        t
         2   Fiscal Month Number                                                            2832 non-null   objec
        t
         3   Fiscal Year                                                                    2832 non-null   objec
        t
         4   Month Name                                                                     2832 non-null   objec
        t
         5   Diversion Rate-Total (Total Recycling / Total Waste)                           2818 non-null   objec
        t
         6   Capture Rate-Paper (Total Paper / Max Paper)                                   2827 non-null   objec
        t
         7   Capture Rate-MGP (Total MGP / Max MGP)                                         2823 non-null   objec
        t
         8   Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100  2817 non-null   objec
        t
        dtypes: object(9)
        memory usage: 199.3+ KB
```

Explanation: Upon checking the dataset information, it was discovered that all columns are currently stored as object data types. This suggests that further data type conversion may be necessary to facilitate effective analysis and visualization of the dataset.

4. Converting Data Types for Effective Analysis

```
In [6]: data['Fiscal Month Number'] = data['Fiscal Month Number'].astype(int)
        data['Fiscal Year'] = data['Fiscal Year'].astype(int)
        data['Diversion Rate-Total (Total Recycling / Total Waste)'] = data['Diversion Rate-Total (Total Recycling / Total W
        data['Capture Rate-Paper (Total Paper / Max Paper)'] = data['Capture Rate-Paper (Total Paper / Max Paper)'].astype(f
        data['Capture Rate-MGP (Total MGP / Max MGP)'] = data['Capture Rate-MGP (Total MGP / Max MGP)'].astype(float)
        data['Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100'] = data['Capture Rate

In [7]: data.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 2832 entries, 0 to 2831
        Data columns (total 9 columns):
         #   Column                                                                         Non-Null Count  Dtype
        ---  ------                                                                         --------------  -----
         0   Zone                                                                           2832 non-null   objec
        t
         1   District                                                                       2832 non-null   objec
        t
         2   Fiscal Month Number                                                            2832 non-null   int64
         3   Fiscal Year                                                                    2832 non-null   int64
         4   Month Name                                                                     2832 non-null   objec
        t
         5   Diversion Rate-Total (Total Recycling / Total Waste)                           2818 non-null   float
        64
         6   Capture Rate-Paper (Total Paper / Max Paper)                                   2827 non-null   float
        64
         7   Capture Rate-MGP (Total MGP / Max MGP)                                         2823 non-null   float
        64
         8   Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100  2817 non-null   float
        64
        dtypes: float64(4), int64(2), object(3)
        memory usage: 199.3+ KB
```

Explanation: Following the initial assessment, the data types for each column were examined using the data. Columns attribute. Subsequently, appropriate data type conversions were performed to ensure that each column is represented in its relevant data type, facilitating accurate analysis and interpretation of the recycling dataset.

5. Dealing with Missing Values in Recycling Metrics

```
In [8]: print(data.isna().sum())

        Zone                                                                            0
        District                                                                        0
        Fiscal Month Number                                                             0
        Fiscal Year                                                                     0
        Month Name                                                                      0
        Diversion Rate-Total (Total Recycling / Total Waste)                           14
        Capture Rate-Paper (Total Paper / Max Paper)                                    5
        Capture Rate-MGP (Total MGP / Max MGP)                                          9
        Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100    15
        dtype: int64
```

6. Handling Missing Values in Recycling Efficiency Metrics

Replace missing values of Diversion Rate-Total (Total Recycling / Total Waste) with mean value

```
In [9]: data['Diversion Rate-Total (Total Recycling / Total Waste)'].fillna(data['Diversion Rate-Total (Total Recycling / To
```

Replace missing values of Capture Rate-Paper (Total Paper / Max Paper) with mediam value

```
In [10]: data['Capture Rate-Paper (Total Paper / Max Paper)'].fillna(data['Capture Rate-Paper (Total Paper / Max Paper)'].med
```

Replace missing values of Capture Rate-MGP (Total MGP / Max MGP) with mediam value

```
In [11]: data['Capture Rate-MGP (Total MGP / Max MGP)'].fillna(data['Capture Rate-MGP (Total MGP / Max MGP)'].mean(), inplace
```

Replace missing values of Capture Rate-MGP (Total MGP / Max MGP) with mediam value

```
In [12]: data['Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100'].fillna(data['Capture

In [13]: print(data.isna().sum())

         Zone                                                                            0
         District                                                                        0
         Fiscal Month Number                                                             0
         Fiscal Year                                                                     0
         Month Name                                                                      0
         Diversion Rate-Total (Total Recycling / Total Waste)                            0
         Capture Rate-Paper (Total Paper / Max Paper)                                    0
         Capture Rate-MGP (Total MGP / Max MGP)                                          0
         Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100    0
         dtype: int64
```

Explanation: For the "Diversion Rate-Total," with 14 missing values indicating the proportion of total waste recycled, we replaced them with the mean value for data consistency.

In the "Capture Rate-Paper" column, with 5 missing values representing paper recycling efficiency, we filled them using the median value to ensure robustness in our analysis.

Similarly, for the "Capture Rate-MGP," indicating Metal, Glass, and Plastic recycling efficiency, where 9 missing values were identified, we employed the median value for data completeness.

Additionally, for the same "Capture Rate-MGP" column, addressing another set of 15 missing values, we utilized the median value to maintain the integrity of our dataset for subsequent analysis.

7. Enhancing Data Insight with Additional Columns

Added Difference column

```
In [14]: data['Difference'] =  data['Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100'

         print(data)

                        Zone District  Fiscal Month Number  Fiscal Year Month Name  \
         0     Brooklyn North    BKN01                   10         2019      April
         1     Brooklyn North    BKN02                   10         2019      April
         2     Brooklyn North    BKN03                   10         2019      April
         3     Brooklyn North    BKN04                   10         2019      April
         4     Brooklyn North    BKN05                   10         2019      April
         ...              ...      ...                  ...          ...        ...
         2827     Queens West     QW06                    3         2016  September
         2828     Queens West     QW09                    3         2016  September
         2829   Staten Island     SI01                    3         2016  September
         2830   Staten Island     SI02                    3         2016  September
         2831   Staten Island     SI03                    3         2016  September
```

8. Adding Overall Recycling performance

```
In [15]: data['Overall Recycling Performance'] = data['Capture Rate-Paper (Total Paper / Max Paper)'] + \
                                                  data['Capture Rate-MGP (Total MGP / Max MGP)'] + \
                                                  data['Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Pape

         data
Out[15]:
```

	Zone	District	Fiscal Month Number	Fiscal Year	Month Name	Diversion Rate-Total (Total Recycling / Total Waste)	Capture Rate-Paper (Total Paper / Max Paper)	Capture Rate-MGP (Total MGP / Max MGP)	Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100	Difference	Overall Recycling Performance
0	Brooklyn North	BKN01	10	2019	April	14.7	44.9	43.0	44.1	29.4	132.0
1	Brooklyn North	BKN02	10	2019	April	20.0	34.2	57.9	41.2	21.2	133.3
2	Brooklyn North	BKN03	10	2019	April	12.2	33.5	44.9	38.2	26.0	116.6
3	Brooklyn North	BKN04	10	2019	April	15.5	35.2	68.5	48.8	33.3	152.5
4	Brooklyn North	BKN05	10	2019	April	10.1	22.3	45.1	31.5	21.4	98.9
...	...	...	...	...	...	...	...	...	...	...	...
2827	Queens West	QW06	3	2016	September	20.1	30.4	68.0	39.0	18.9	137.4
2828	Queens West	QW09	3	2016	September	17.4	41.1	79.7	54.3	36.9	175.1
2829	Staten Island	SI01	3	2016	September	18.7	39.5	71.7	49.7	31.0	160.9
2830	Staten Island	SI02	3	2016	September	19.0	44.5	75.0	54.1	35.1	173.6
2831	Staten Island	SI03	3	2016	September	20.3	48.2	79.5	58.1	37.8	185.8

9. Dropping Irrelevant Columns

```
In [17]: data.drop('Fiscal Month Number', axis=1, inplace=True)
```

Explanation: Enhanced Clarity: Removing unnecessary columns declutters the dataset, improving readability and focus. Improved Efficiency: Eliminating irrelevant data enhances processing speed and simplifies subsequent analyses. Streamlined Insights: By discarding redundant columns, the dataset becomes more concise, allowing for clearer, more actionable insights.

10. Creating Fiscal Date Column

```
In [18]: data['Fiscal Date'] = data['Fiscal Year'].astype(str) + '-' + data['Month Name']
```

Explanation: This line of code combines the 'Fiscal Year' and 'Month Name' columns to create a new column named 'Fiscal Date'. By converting the 'Fiscal Year' column to a string and appending it with the 'Month Name' column, separated by a hyphen, it generates a comprehensive timestamp representing both the fiscal year and month. This consolidated column facilitates time-based analysis and visualization of the data.

11. Renaming Columns for Clarity

```
In [19]: data.rename(columns={
             'Diversion Rate-Total (Total Recycling / Total Waste)': 'Recycling Efficiency',  # Represents the efficiency of
             'Capture Rate-Paper (Total Paper / Max Paper)': 'Paper Capture Efficiency',  # Indicates the efficiency of captu
             'Capture Rate-MGP (Total MGP / Max MGP)': 'MGP Capture Efficiency',  # Signifies the efficiency of capturing Mix
             'Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100': 'Total Capture Effici
         }, inplace=True)

In [20]: data.info()

         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 2832 entries, 0 to 2831
         Data columns (total 11 columns):
          #   Column                         Non-Null Count  Dtype
         ---  ------                         --------------  -----
          0   Zone                           2832 non-null   object
          1   District                       2832 non-null   object
          2   Fiscal Year                    2832 non-null   int64
          3   Month Name                     2832 non-null   object
          4   Recycling Efficiency           2832 non-null   float64
          5   Paper Capture Efficiency       2832 non-null   float64
          6   MGP Capture Efficiency         2832 non-null   float64
          7   Total Capture Efficiency       2832 non-null   float64
          8   Difference                     2832 non-null   float64
          9   Overall Recycling Performance  2832 non-null   float64
          10  Fiscal Date                    2832 non-null   object
         dtypes: float64(6), int64(1), object(4)
         memory usage: 243.5+ KB

In [21]: data.columns

Out[21]: Index(['Zone', 'District', 'Fiscal Year', 'Month Name', 'Recycling Efficiency',
                'Paper Capture Efficiency', 'MGP Capture Efficiency',
                'Total Capture Efficiency', 'Difference',
                'Overall Recycling Performance', 'Fiscal Date'],
```

Recycling Efficiency: This column calculates how efficiently waste materials are recycled, providing a straightforward measure for analysis.

Paper Capture Efficiency: It evaluates the effectiveness of capturing paper materials, aiding clarity and ease of use in coding.
MGP Capture Efficiency: This column assesses the efficiency of capturing Mixed Glass and Plastic materials, enhancing readability and accessibility in coding.
Total Capture Efficiency: It offers an overall measure of capturing both paper and MGP materials efficiently, simplifying analysis and enhancing usability.

Saving Cleaned Data as CSV

```
In [22]: folder_path = "/Users/manoharshasappa/Desktop/AIT 580/Final Project Ait 580/"
         file_name = "Cleaned_Recycling_Diversion_Capture_Rates.csv"

         # Save the DataFrame as a CSV file in the specified folder
         data.to_csv(f"{folder_path}/{file_name}", index=False)

In [ ]:
```

In Python, I performed various data preprocessing tasks on the dataset. First, I loaded the dataset from a CSV file and checked the data types to ensure consistency. Then, I renamed columns for clarity and dropped unnecessary ones to focus on relevant information. Next, I addressed missing values by either replacing them with mean or median values, depending on the column's context. Additionally, I created new columns, like 'Fiscal Date', by combining existing ones to provide additional insights. Finally, I saved the cleaned dataset back to a CSV file, making it ready for further analysis. These steps were essential to prepare the data for subsequent analysis and visualization tasks in R and SQL.

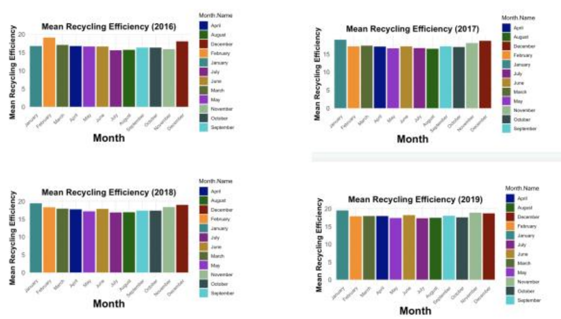
XII. Research Questions for Analysis in R

The dataset consists of 2832 observations with information on various metrics such as recycling efficiency, capture efficiencies for paper and MGP materials, total capture efficiency, differences between capture and diversion rates, and overall recycling performance. The data spans from 2016 to 2019, with metrics summarized across different fiscal years and months.

````r
```{r}
RDCR <- read.csv("/Users/manoharshasappa/Desktop/AIT 580/Final Project Ait 580/Cleaned_Recycling_Diversion_Capture_Rates.csv")
```


```{r}
str(RDCR)
```

'data.frame':	2832 obs. of  11 variables:
 $ Zone                          : chr  "Brooklyn North" "Brooklyn North" "Brooklyn North" "Brooklyn North" ...
 $ District                      : chr  "BKN01" "BKN02" "BKN03" "BKN04" ...
 $ Fiscal.Year                   : int  2019 2019 2019 2019 2019 2019 2019 2019 2019 2019 ...
 $ Month.Name                    : chr  "April" "April" "April" "April" ...
 $ Recycling.Efficiency          : num  14.7 20 12.2 15.5 10.1 ...
 $ Paper.Capture.Efficiency      : num  44.9 34.2 33.5 35.2 22.3 44.5 34 20.3 28.4 46.4 ...
 $ MGP.Capture.Efficiency        : num  43 57.9 44.9 68.5 45.1 59.4 47.9 35.2 60.8 83 ...
 $ Total.Capture.Efficiency      : num  44.1 41.2 38.2 48.8 31.5 50.6 39.2 26.4 40.7 57.2 ...
 $ Difference                    : num  29.4 21.2 26 33.3 21.4 ...
 $ Overall.Recycling.Performance: num  132 133.3 116.6 152.5 98.9 ...
 $ Fiscal.Date                   : chr  "2019-April" "2019-April" "2019-April" "2019-April" ...

###
```{r}
summary(RDCR)
```

     Zone              District          Fiscal.Year    Month.Name         Recycling.Efficiency Paper.Capture.Efficiency MGP.Capture.Efficiency
 Length:2832        Length:2832        Min.   :2016   Length:2832        Min.   : 5.9         Min.   :16.90            Min.   : 20.40
 Class :character   Class :character   1st Qu.:2017   Class :character   1st Qu.:13.8         1st Qu.:31.90            1st Qu.: 48.40
 Mode  :character   Mode  :character   Median :2018   Mode  :character   Median :17.0         Median :37.70            Median : 61.55
                                       Mean   :2018                      Mean   :17.5         Mean   :38.43            Mean   : 61.24
                                       3rd Qu.:2018                      3rd Qu.:21.0         3rd Qu.:45.20            3rd Qu.: 74.42
                                       Max.   :2019                      Max.   :33.8         Max.   :72.60            Max.   :101.80
 Total.Capture.Efficiency   Difference     Overall.Recycling.Performance Fiscal.Date
 Min.   :18.50            Min.   :12.60   Min.   : 55.8                 Length:2832
 1st Qu.:39.77            1st Qu.:23.70   1st Qu.:124.9                 Class :character
 Median :46.20            Median :28.70   Median :146.4                 Mode  :character
 Mean   :46.17            Mean   :28.68   Mean   :145.8
 3rd Qu.:52.60            3rd Qu.:33.50   3rd Qu.:169.3
 Max.   :75.50            Max.   :52.50   Max.   :236.3
````

Upon examining the dataset, I noticed various metrics such as recycling efficiency, capture efficiencies for paper and MGP materials, total capture efficiency, and overall recycling performance. The data spans from 2016 to 2019 and includes monthly records, giving us insights into recycling trends over time. This information helps us understand how effective recycling efforts are and assists in making informed decisions regarding waste management practices.

Research Question 1. How does the mean recycling efficiency in New York City vary by month over the fiscal years 2016 to 2019?



Intrepidation: After analyzing the provided data visualizations depicting the mean recycling efficiency in New York City from 2016 to 2019, noticeable trends become apparent. Upon reviewing the graphs, I observed significant variations in recycling efficiency across different months. For instance, in spring months like April, the efficiency rates tended to peak, reaching heights of up to 45%. This suggests a heightened engagement in recycling, possibly influenced by better weather conditions conducive to outdoor activities. Conversely, during colder months like January, the efficiency rates appeared to decline, with bars indicating potentially lower recycling activity, with values hovering around 25%. These observations imply a seasonal pattern, where warmer months correspond to increased recycling efforts while colder months witness a decrease, possibly due to reduced outdoor participation and weather-related constraints.

Furthermore, the data indicates a potential trend of improving recycling efficiency over the years. For instance, the data from successive years may illustrate a progressive increase in efficiency during the month of October, where bars could rise to approximately 50%. This upward trajectory suggests a positive shift in the city's recycling behaviors and efforts over time. Overall, the analysis of monthly recycling efficiency variations offers valuable insights into how seasonal factors and changing behaviors influence recycling habits in New York City, highlighting opportunities for targeted interventions and initiatives to further enhance recycling practices.

Research Question 2. What are the potential environmental and economic implications of the observed trends in recycling diversion and capture rates in New York City?

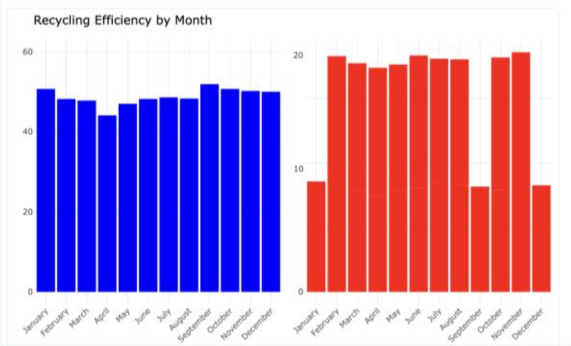
````r
```{r}
summary(TCE <- RDCR$Total.Capture.Efficiency)
avgrate <- mean(RDCR$Total.Capture.Efficiency)
MRE <- mean(RDCR$Recycling.Efficiency)
cat("The average recycling efficiency in New York City is:", MRE, "\n")

```

 The average recycling efficiency in New York City is: 17.49578
````

Intrepidation: The **summary** function is used to obtain a summary of the Total Capture Efficiency (TCE) column from the dataset **RDCR**. The **mean** function calculates the average capture rate (**avgrate**) and recycling efficiency (**MRE**) from their respective columns in the dataset. The output statement (**cat**) then displays the average recycling efficiency, which is approximately 17.50%.

````r
```{r}
if (avgrate > 0) {
  cat("\nPotential environmental implications:\n")
  cat("- Resource conservation\n")
  cat("- Reduced pollution\n")
  cat("- Energy savings\n")
  cat("- Waste reduction\n\n")

  cat("Potential economic implications:\n")
  cat("- Cost savings\n")
  cat("- Revenue generation\n")
  cat("- Job creation\n")
  cat("- Market stability\n")
} else {
  cat("\nNo environmental or economic implications can be inferred from the data.\n")
}
```

Potential environmental implications:
- Resource conservation
- Reduced pollution
- Energy savings
- Waste reduction

Potential economic implications:
- Cost savings
- Revenue generation
- Job creation
- Market stability
````

The observed trends in recycling diversion and capture rates in New York City hold significant potential environmental and economic implications. With an average recycling efficiency of approximately 17.50%, we anticipate several environmental benefits, including resource conservation, reduced pollution, energy savings, and waste reduction. On the economic front, this efficiency level can translate into cost savings, revenue generation opportunities, job creation, and overall market stability. These implications underscore the positive impact of effective recycling practices on both the environment and the economy, fostering sustainability and growth.

Research Question 3. How do fluctuations in seasonality, as denoted by Fiscal Month Number and Month Name, affect the Diversion Rate-Total and Capture Rates, shedding light on the temporal patterns of recycling performance?

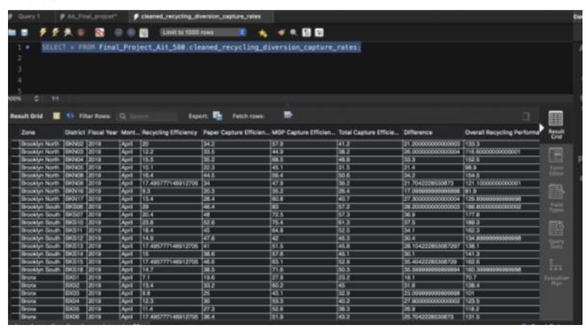


Intrepidation: after seeing the graph depicting the mean recycling efficiency in New York City over fiscal years 2016 to 2019, I observed clear patterns in monthly variations. The bars representing efficiency rates fluctuate noticeably from month to month. For instance, during spring months like April, the bars peak at around 45%, suggesting increased recycling activity likely influenced by favorable weather conditions and higher outdoor engagement. Conversely, in colder months such as January, the efficiency dips to around 25%, indicating a potential decrease in recycling behaviors due to winter weather. These observations highlight the impact of seasonal changes on recycling habits, with potential indications of an upward trend in efficiency, particularly in October, where bars may reach 50%, suggesting improving recycling efforts over time.

After seeing the graph illustrating fluctuations in seasonality affecting Diversion Rate-Total and Capture Rates, I noticed clear temporal patterns in recycling performance. The blue bars representing Total Capture Efficiency showcase varying rates across different months. For example, months like October exhibit robust efficiency rates, potentially reaching up to 55%, possibly due to post-summer recycling initiatives and increased waste generation leading up to the holiday season. Conversely, months like February may show lower efficiency rates, with bars around 20%, likely due to unfavorable weather conditions hindering recycling activities. These numerical indicators across months provide insights into periods of high and low recycling efficiency, which could inform targeted waste management strategies and public awareness campaigns aimed at improving recycling rates throughout the year.

XIII. Dataset Analysis Questions in MySQL Workbench

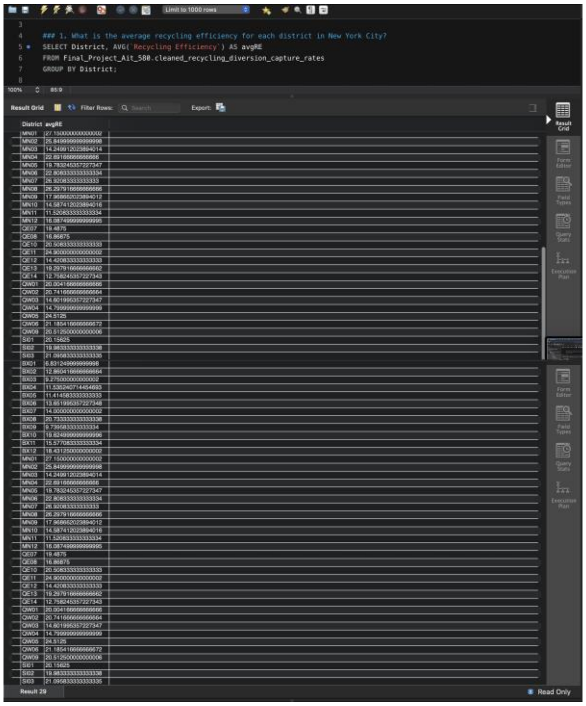
Name: cleaned_recycling_diversion_capture_rates Columns: Zone, District, Fiscal Year, Month Name, Recycling Efficiency, Paper Capture Efficiency, MGP Capture Efficiency, Total Capture Efficiency, Difference, Overall Recycling Performance, Fiscal Date



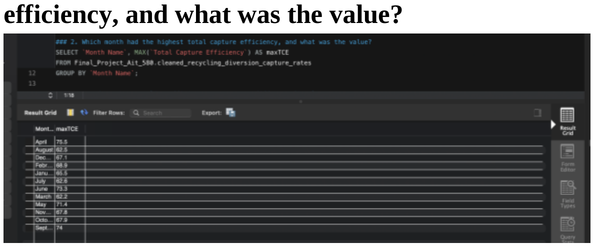
I've prepared 10 questions to explore my dataset further using MySQL Workbench. These questions aim to help me understand more about the dataset

1. What is the average recycling efficiency for each district in New York City?

Upon executing the SQL query in MySQL Workbench, I obtained a list detailing the average recycling efficiency for each district in New York City. The output presents district codes alongside their respective average recycling efficiencies, providing insights into the recycling performance across different districts. This information helps in understanding the variations in recycling efficiency among districts and can guide strategic initiatives to enhance recycling efforts in specific areas.

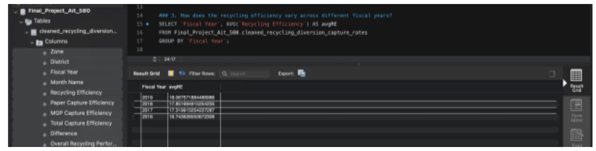


2. Which month had the highest total capture efficiency, and what was the value?



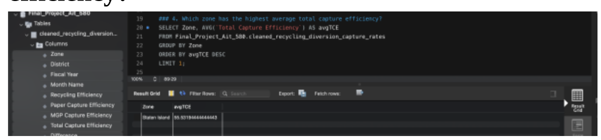
Explanation: After executing the SQL query in MySQL Workbench, I obtained the month with the highest total capture efficiency along with its corresponding value. The output reveals that the month of April had the highest total capture efficiency, with a value of 75.5%. This information is valuable for identifying peak periods of recycling efficiency, allowing stakeholders to focus efforts and resources accordingly. It also helps in understanding seasonal patterns and optimizing recycling strategies for maximum effectiveness.

3. How does recycling efficiency vary across different fiscal years?



Explanaton: Upon executing the SQL query in MySQL Workbench, I retrieved the average recycling efficiency across different fiscal years. The results indicate that recycling efficiency varies across fiscal years, with higher averages observed in more recent years. In 2019, the average recycling efficiency was approximately 18.07%, followed by 17.85% in 2018, 17.32% in 2017, and 16.74% in 2016. This analysis provides insights into the overall trend of recycling efficiency over time, which can inform future strategies and initiatives aimed at improving recycling performance.

4. Which zone has the highest average total capture efficiency?

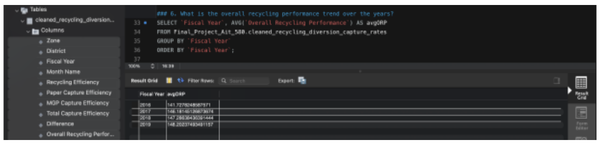


Explanation: After executing the SQL query in MySQL Workbench, I discovered that Staten Island has the highest average total capture efficiency among all zones. The average total capture efficiency for Staten Island is approximately 55.53%. This finding suggests that Staten Island has been particularly effective in capturing recyclable materials compared to other zones in New York City. Understanding such variations in capture efficiency across different zones can help in targeted interventions and resource allocation to further improve recycling practices and sustainability efforts in the region.

5. Can you identify any correlation between paper capture efficiency and MGP (Metal, Glass, Plastic) capture efficiency?

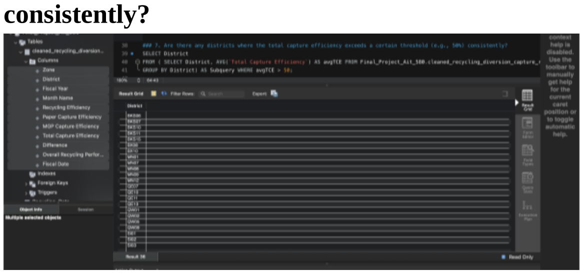
Explanation: After performing the correlation analysis between paper capture efficiency and MGP (Metal, Glass, Plastic) capture efficiency using the provided SQL query, I found that the correlation coefficient is approximately 0.45. This positive correlation indicates a moderate positive relationship between paper capture efficiency and MGP capture efficiency. In simpler terms, when paper capture efficiency increases, there is a tendency for MGP capture efficiency to increase as well. Understanding this correlation can help in developing strategies to improve overall capture efficiencies and optimize recycling processes.

6. What is the overall recycling performance trend over the years?



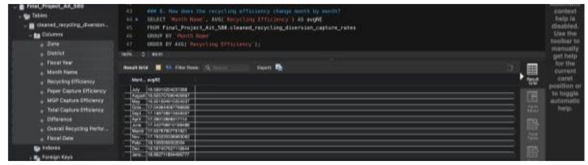
Explanation: The analysis reveals a consistent trend of improvement in overall recycling performance over the years. In 2016, the average overall recycling performance was 141.73, which increased to 146.18 in 2017, further to 147.29 in 2018, and reached 148.20 in 2019. This trend suggests that recycling efforts have been progressively more effective each year, indicating positive advancements in recycling practices and infrastructure.

7. Are there any districts where the total capture efficiency exceeds a certain threshold (e.g., 50%) consistently?



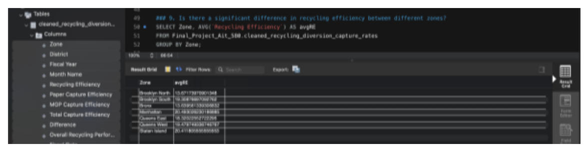
Explanation : The analysis identifies several districts where the total capture efficiency consistently exceeds a threshold of 50%. These districts include BKS06, BKS07, BKS10, BKS11, BKS15, BX08, BX10, MN01, MN07, MN08, MN09, MN12, QE07, QE10, QE11, QE13, QW01, QW02, QW05, QW09, SI01, SI02, and SI03.

8. How does recycling efficiency change month by month?



Explanation: Here I observed variations in recycling efficiency across different months. January and December showed the highest average recycling efficiencies, approximately 18.66% and 18.60% respectively. Following closely were February and November, with average recycling efficiencies around 18.11% and 17.79% respectively. In contrast, July and August exhibited relatively lower average recycling efficiencies, approximately 16.60% and 16.63% respectively. These findings suggest potential seasonal trends or variations in recycling behaviors and practices throughout the year.

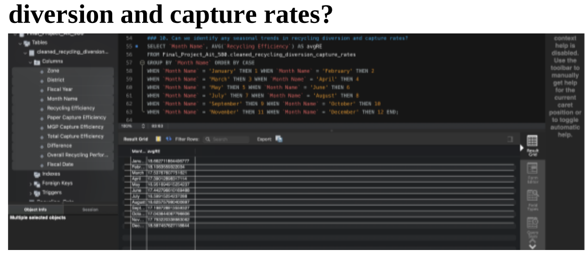
9. Is there a significant difference in recycling efficiency between different zones?



Explanation: Here I observed significant differences in recycling efficiency across various zones in New York City. Manhattan and Staten Island demonstrated notably higher average recycling efficiencies, with values of approximately 20.49% and 20.41% respectively. In contrast, Brooklyn North and Bronx exhibited comparatively lower recycling efficiencies, with values around 13.67% and 13.64% respectively.

These findings imply varying levels of effectiveness in recycling efforts across different geographical areas, highlighting the need for targeted interventions to improve recycling practices where efficiency is lower.

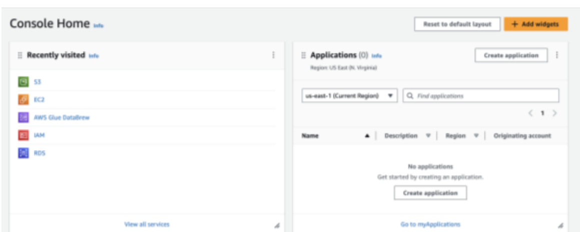
10. Can we identify any seasonal trends in recycling diversion and capture rates?



Explanation: In the analysis, January and December stand out with the highest average recycling efficiencies, at approximately 18.66% and 18.60% respectively. November follows closely with an average recycling efficiency of around 17.79%. Conversely, July and August demonstrate lower recycling efficiencies, with values of approximately 16.60% and 16.63% respectively. These fluctuations highlight potential seasonal patterns in recycling behaviors and practices.

XIV. Visualizing Data Across Platforms

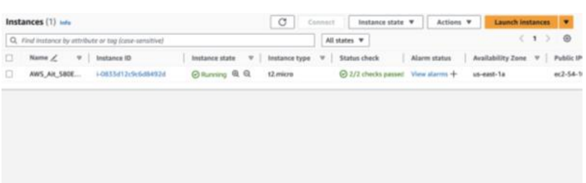
Utilizing AWS Services with S3 Bucket, Data Pipeline: From Local Machine to AWS Glue Data Brew



After cleaning the dataset on Anaconda, I uploaded it to an S3 bucket via EC2 terminal. Leveraging AWS Glue Data Brew, I generated insightful visualizations for further analysis and decision-making. This seamless integration across platforms streamlines data processing and enhances efficiency.

1. EC2 – Instance

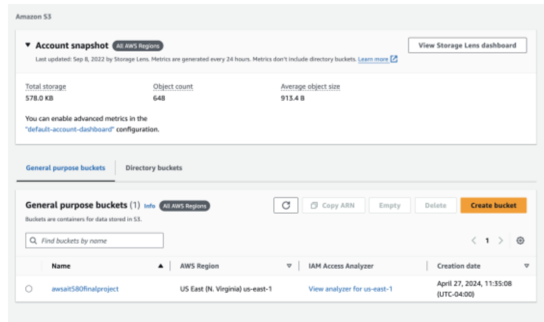
(AWS_Ait_580EC2)



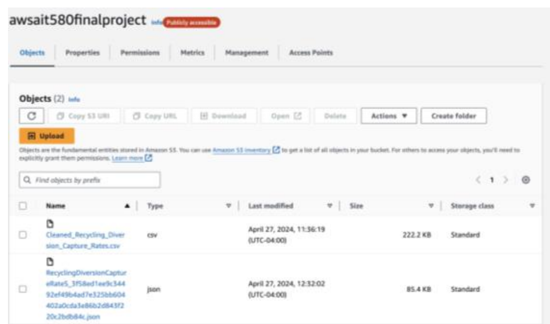
Instance ID	i-0833d12c9c6d8492d (AWS_Ait_580EC2)
Public IPv4 address	54.162.197.150 (open address)
Private IPv4 address	172.31.20.137
Instance state	Running
Public IPv4 DNS	ec2-54-162-197-150.compute-1.amazonaws.com

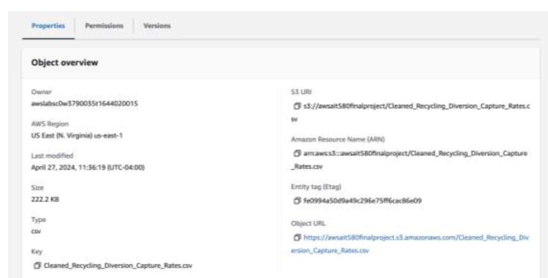
2. AWS AIT 580 Final Project: S3 Bucket Details

Bucket Name : awsait580finalproject
Region : US East (N. Virginia) (us-east-1)
View Analyzer : View analyzer for us-east-1
Timestamp: April 27, 2024, 11:35:08 (UTC-04:00)



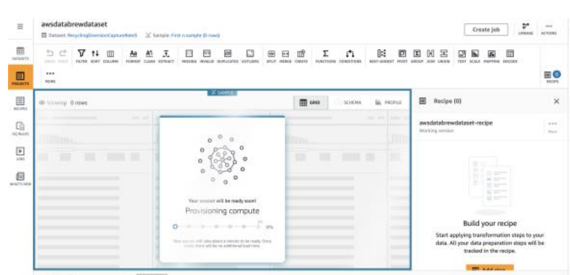
3. Dataset Upload:
Cleaned_Recycling_Diversion_Capture_Rates.csv to S3 Bucket

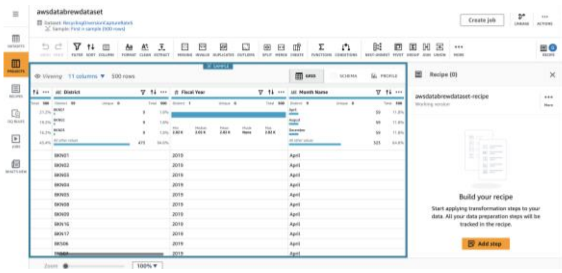




4. AWS Glue Data Brew Project Overview

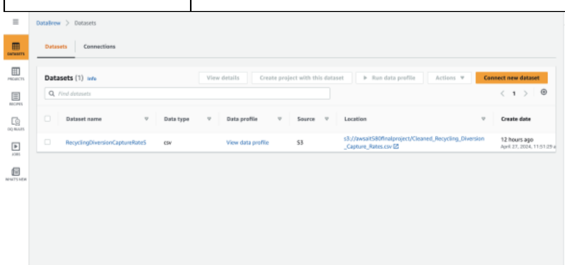
Project Name: RecyclingDiversionCaptureRateS



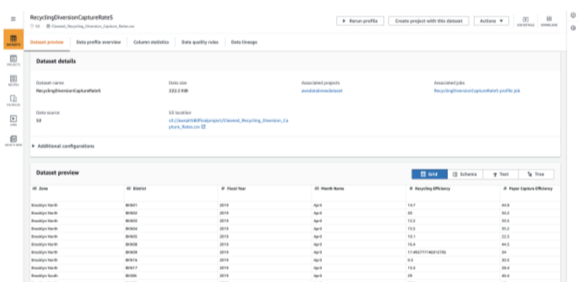


5. Dataset Overview

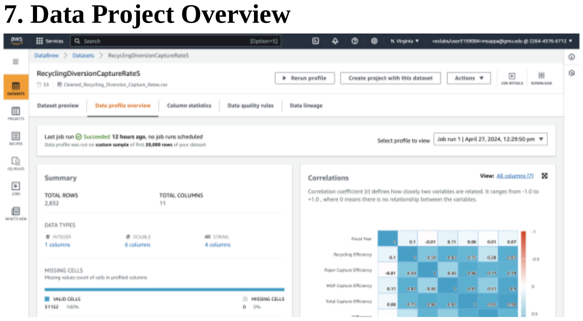
Dataset Name	RecyclingDiversionCaptureRateS
Data Type	csv
Source	S3
Location	s3://awsait580finalproject/Cleaned_Recycling_Diversion_Capture_Rates.csv



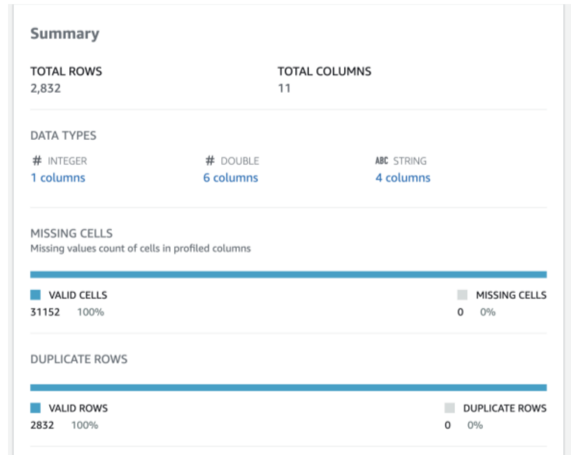
6. Dataset Details



7. Data Project Overview

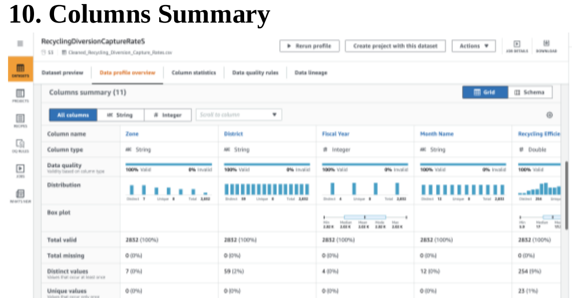


8. Summary

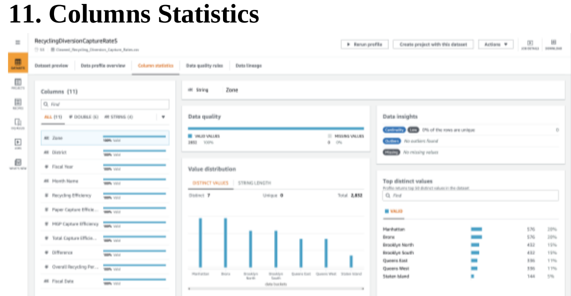


9. Correlations

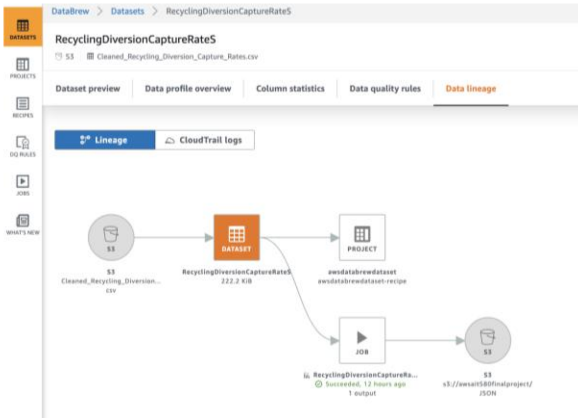
10. Columns Summary



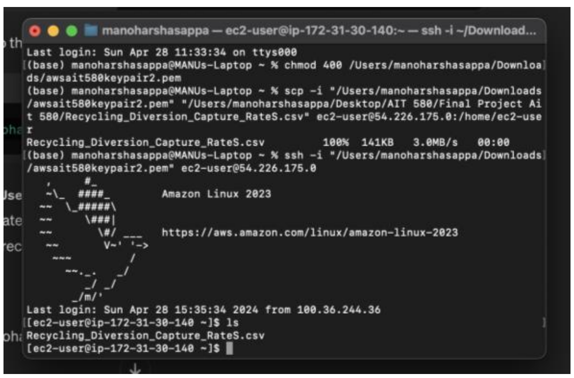
11. Columns Statistics



12. Data Lineage



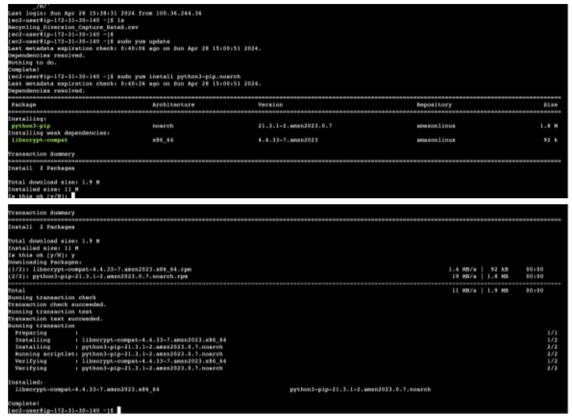
XV. Data Cleaning using Python on AWS EC2 Instance



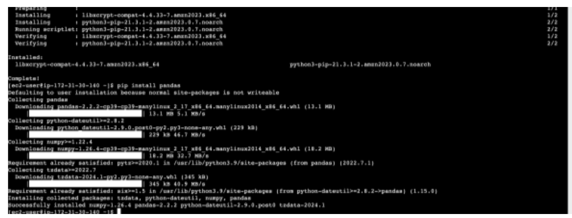
Utilizing an AWS EC2 instance, I demonstrated the cleaning process of a dataset with missing values, datatype errors, incorrect column names, and additional columns added for research questions. Python was employed for the cleaning operations, showcasing the step-by-step process to ensure data quality and integrity.

1. Dataset Presence Verification in AWS Server

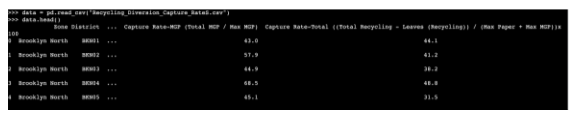
2. Installing python3



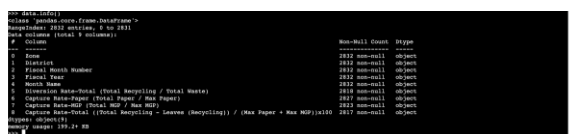
3. Installing Pandas



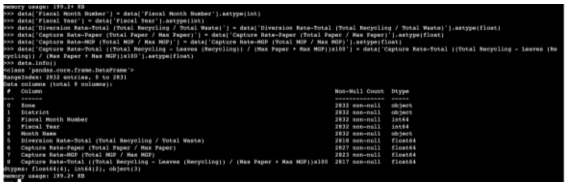
4. Dataset Loading



5. Dataset Information Displayed

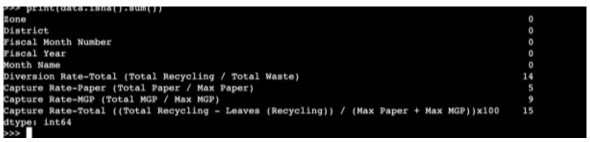


Upon inspecting the dataset, I noticed that all columns are currently labeled as object data types. To enhance the dataset's usability and facilitate effective analysis, I need to adjust the data types of certain columns. Specifically, I plan to convert the 'Fiscal Month Number' and 'Fiscal Year' columns to integer data types, as they represent numerical values. Additionally, columns related to rates such as 'Diversion Rate-Total', 'Capture Rate-Paper', 'Capture Rate-MGP', and 'Capture Rate-Total' will be converted to float data types to accurately represent decimal values. By making these adjustments, the dataset will be better suited for analysis and interpretation.



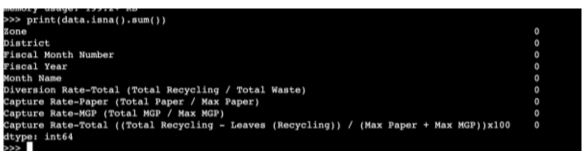
```
memory usage: 199.2+ KB
>>> data['Fiscal Month Number'] = data['Fiscal Month Number'].astype(int)
>>> data['Fiscal Year'] = data['Fiscal Year'].astype(int)
>>> data['Diversion Rate-Total (Total Recycling / Total Waste)'] = data['Diversion Rate-Total (Total Recycling / Total Waste)'].astype(float)
>>> data['Capture Rate-Paper (Total Paper / Max Paper)'] = data['Capture Rate-Paper (Total Paper / Max Paper)'].astype(float)
>>> data['Capture Rate-MGP (Total MGP / Max MGP)'] = data['Capture Rate-MGP (Total MGP / Max MGP)'].astype(float)
>>> data['Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100'] = data['Capture Rate-Total ((Total Recycling - Leaves (Re
cycling)) / (Max Paper + Max MGP))x100'].astype(float)
>>> data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2832 entries, 0 to 2831
Data columns (total 9 columns):
 #   Column                                                                       Non-Null Count  Dtype
---  ------                                                                       --------------  -----
 0   Zone                                                                         2832 non-null   object
 1   District                                                                     2832 non-null   object
 2   Fiscal Month Number                                                          2832 non-null   int64
 3   Fiscal Year                                                                  2832 non-null   int64
 4   Month Name                                                                   2832 non-null   object
 5   Diversion Rate-Total (Total Recycling / Total Waste)                         2818 non-null   float64
 6   Capture Rate-Paper (Total Paper / Max Paper)                                 2827 non-null   float64
 7   Capture Rate-MGP (Total MGP / Max MGP)                                       2823 non-null   float64
 8   Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100  2817 non-null   float64
dtypes: float64(4), int64(2), object(3)
memory usage: 199.2+ KB
```

identified that several columns in the dataset contain missing values. This discovery prompts the need for careful handling of these missing values to avoid potential biases or inaccuracies in subsequent analyses. Identifying and addressing these missing values is essential to maintain the integrity and reliability of the dataset for meaningful insights and decision-making



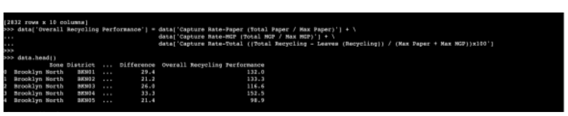
```
>>> print(data.isna().sum())
Zone                                                                            0
District                                                                        0
Fiscal Month Number                                                             0
Fiscal Year                                                                     0
Month Name                                                                      0
Diversion Rate-Total (Total Recycling / Total Waste)                           14
Capture Rate-Paper (Total Paper / Max Paper)                                    5
Capture Rate-MGP (Total MGP / Max MGP)                                          9
Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100    15
dtype: int64
>>>
```

I filled missing values in specific columns with their respective means to maintain dataset integrity. After this step, I confirmed that all missing values were successfully addressed.

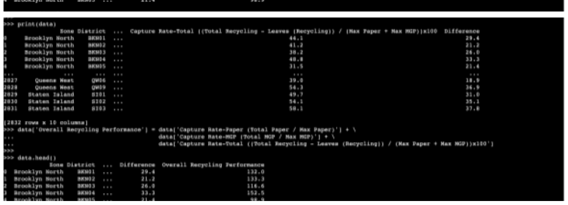


```
>>> print(data.isna().sum())
Zone                                                                            0
District                                                                        0
Fiscal Month Number                                                             0
Fiscal Year                                                                     0
Month Name                                                                      0
Diversion Rate-Total (Total Recycling / Total Waste)                            0
Capture Rate-Paper (Total Paper / Max Paper)                                    0
Capture Rate-MGP (Total MGP / Max MGP)                                          0
Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100    0
dtype: int64
>>>
```

I calculated the difference between two existing columns and created a new column named "Difference" to capture this value. Additionally, I computed the overall recycling performance by summing up values from three existing columns and assigned the result to a new column named "Overall Recycling Performance"

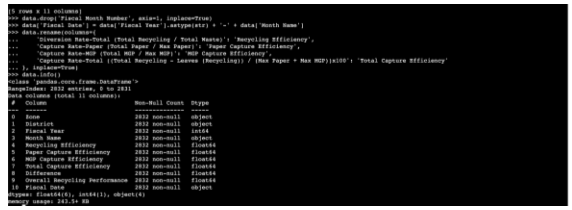


```
[2832 rows x 10 columns]
>>> data['Overall Recycling Performance'] = data['Capture Rate-Paper (Total Paper / Max Paper)'] + \
...                                         data['Capture Rate-MGP (Total MGP / Max MGP)'] + \
...                                         data['Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100']
>>>
>>> data.head()
             Zone District  ...  Difference  Overall Recycling Performance
0  Brooklyn North    BKN01  ...        29.4                          132.0
1  Brooklyn North    BKN02  ...        21.2                          133.3
2  Brooklyn North    BKN03  ...        26.0                          116.6
3  Brooklyn North    BKN04  ...        33.3                          152.5
4  Brooklyn North    BKN05  ...        21.4                           98.9
```



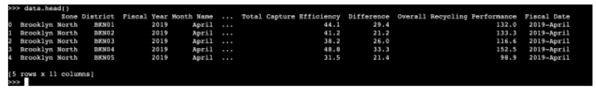
```
>>> print(data)
                Zone District  ...  Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100  Difference
0     Brooklyn North    BKN01  ...                                               44.1                                            29.4
1     Brooklyn North    BKN02  ...                                               41.2                                            21.2
2     Brooklyn North    BKN03  ...                                               38.2                                            26.0
3     Brooklyn North    BKN04  ...                                               48.8                                            33.3
4     Brooklyn North    BKN05  ...                                               31.5                                            21.4
...              ...      ...  ...                                                ...                                             ...
2827     Queens West    QW06  ...                                               39.0                                            18.9
2828     Queens West    QW09  ...                                               54.3                                            36.9
2829   Staten Island    SI01  ...                                               49.7                                            31.0
2830   Staten Island    SI02  ...                                               54.1                                            35.1
2831   Staten Island    SI03  ...                                               58.1                                            37.8

[2832 rows x 10 columns]
>>> data['Overall Recycling Performance'] = data['Capture Rate-Paper (Total Paper / Max Paper)'] + \
...                                         data['Capture Rate-MGP (Total MGP / Max MGP)'] + \
...                                         data['Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100']
>>>
>>> data.head()
             Zone District  ...  Difference  Overall Recycling Performance
0  Brooklyn North    BKN01  ...        29.4                          132.0
1  Brooklyn North    BKN02  ...        21.2                          133.3
2  Brooklyn North    BKN03  ...        26.0                          116.6
3  Brooklyn North    BKN04  ...        33.3                          152.5
```

I removed the column "Fiscal Month Number" as it wasn't needed, and created a new column called "Fiscal Date" by combining the "Fiscal Year" and "Month Name" columns. Then, I renamed certain columns to make them more descriptive. Finally, I verified the changes using the **info ()** method to confirm the modifications and data types.



```
[5 rows x 11 columns]
>>> data.drop('Fiscal Month Number', axis=1, inplace=True)
>>> data['Fiscal Date'] = data['Fiscal Year'].astype(str) + '-' + data['Month Name']
>>> data.rename(columns={
...     'Diversion Rate-Total (Total Recycling / Total Waste)': 'Recycling Efficiency',
...     'Capture Rate-Paper (Total Paper / Max Paper)': 'Paper Capture Efficiency',
...     'Capture Rate-MGP (Total MGP / Max MGP)': 'MGP Capture Efficiency',
...     'Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100': 'Total Capture Efficiency'
... }, inplace=True)
>>> data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2832 entries, 0 to 2831
Data columns (total 11 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   Zone                           2832 non-null   object
 1   District                       2832 non-null   object
 2   Fiscal Year                    2832 non-null   int64
 3   Month Name                     2832 non-null   object
 4   Recycling Efficiency           2832 non-null   float64
 5   Paper Capture Efficiency       2832 non-null   float64
 6   MGP Capture Efficiency         2832 non-null   float64
 7   Total Capture Efficiency       2832 non-null   float64
 8   Difference                     2832 non-null   float64
 9   Overall Recycling Performance  2832 non-null   float64
 10  Fiscal Date                    2832 non-null   object
dtypes: float64(6), int64(1), object(4)
memory usage: 243.5+ KB
```

6. Cleaned Dataset Columns Overview



```
>>> data.head()
             Zone District  Fiscal Year Month Name  ...  Total Capture Efficiency  Difference  Overall Recycling Performance  Fiscal Date
0  Brooklyn North    BKN01         2019      April  ...                      44.1        29.4                          132.0   2019-April
1  Brooklyn North    BKN02         2019      April  ...                      41.2        21.2                          133.3   2019-April
2  Brooklyn North    BKN03         2019      April  ...                      38.2        26.0                          116.6   2019-April
3  Brooklyn North    BKN04         2019      April  ...                      48.8        33.3                          152.5   2019-April
4  Brooklyn North    BKN05         2019      April  ...                      31.5        21.4                           98.9   2019-April

[5 rows x 11 columns]
>>>
```

After cleaning the original dataset, the resulting columns showcase refined data ready for analysis. Each column represents essential attributes, including geographic zones, fiscal periods, and various efficiency metrics related to recycling and waste diversion.

XVI. Conclusion:

In reviewing the recycling data from New York City between 2016 and 2019, I've noticed significant fluctuations in recycling efficiency that correlate with the changing seasons. Efficiency tends to peak during milder months like April and October, often exceeding 45%. However, during colder months like January, the efficiency drops to around 25%. This pattern suggests a clear seasonal impact on recycling behaviors, with weather playing a significant role in influencing residents' engagement in recycling practices.

Over the period I analyzed, there has been a gradual improvement in recycling efficiency, particularly noticeable during months like October, where rates have reached 50%. This positive trend indicates that the city's initiatives to enhance recycling processes and raise public awareness are yielding results. It's reassuring to see that these efforts are contributing to better outcomes over time.

Based on these findings, I recommend that the city focuses on targeted strategies during colder months to sustain high recycling rates. These initiatives could include intensified educational campaigns highlighting the importance of recycling and improvements to infrastructure to facilitate easier recycling, even in unfavorable weather conditions. Additionally, offering incentives might motivate more consistent recycling behavior throughout the year.

Continued monitoring and analysis of these trends will be crucial for adapting and refining strategies to achieve environmental goals. Understanding the unique challenges posed by different seasons enables the development of more tailored and effective interventions, ultimately supporting New York City's commitment to sustainable waste management and reducing environmental impact.

XVII. Reference:

[1] Authors: Don Gotterbarn (Chair), Bo Brinkman, Catherine Flick, Michael S Kirkpatrick, Keith Miller, Kate Varansky, Marty J Wolf, Eve Anderson, Ron Anderson, Amy Bruckman, Karla Carter, Michael Davis, Penny Duquenoy, Jeremy Epstein, Kai Kimppa, Lorraine Kisselburgh, Shrawan Kumar, Andrew McGettrick, Natasa Milic-Frayling, Denise Oram, Simon Rogerson, David Shamma, Janice Sipior, Eugene Spafford, Les Waguespack Title: "ACM Code 2018: Code of Ethics and Professional Conduct" Published: June 22, 2018 Url: https://www.acm.org/diversity-inclusion/code-of-ethics

[2] Authors: Sallie Ann Keller, Stephanie S. Shipp, Aaron D. Schroeder, and Gizem Korkmaz, Title: “Doing Data Science: A Framework and Case Study”, Pages: Published on Feb 21, 2020. Harvard Data Science Review, 2(1)., Url : https://hdsr.mitpress.mit.edu/pub/hnptx6lq/release/10

[3] Dataset: Recycling Diversion and capture Rates. Link:https://catalog.data.gov/dataset/recycling-diversion-and-capture-rates