Machine Learning Project 1

Jonathan Chacko Pattasseril

Marine Data Analysis

Report

Table of Contents

[Comprehensive Report on Marine Data Analysis 2](#_Toc190278372)

[Dataset Citation: 2](#_Toc190278373)

[1. Dataset Selection and Preliminary Research 3](#_Toc190278374)

[Goals and Questions 3](#_Toc190278375)

[Steps Performed 3](#_Toc190278376)

[ Dataset Selection: 3](#_Toc190278377)

[ Preliminary Research: 3](#_Toc190278378)

[Important Outputs: 4](#_Toc190278379)

[2. Data Cleaning and Preprocessing 5](#_Toc190278380)

[Objectives 5](#_Toc190278381)

[Steps Performed 5](#_Toc190278382)

[a) Environment Setup and Data Retrieval: 5](#_Toc190278383)

[b) Initial Data Exploration: 6](#_Toc190278384)

[c) Date Cleaning and Preprocessing: 6](#_Toc190278385)

[d) Data Standardisation: 6](#_Toc190278386)

[e) Final Preprocessing Steps: 6](#_Toc190278387)

[3. Exploratory Data Analysis (EDA) 7](#_Toc190278388)

[Objectives 7](#_Toc190278389)

[Steps Performed 7](#_Toc190278390)

[a) Data Filtering for Analysis: 7](#_Toc190278391)

[b) Heatmaps and Pivot Tables: 7](#_Toc190278392)

[4. Statistical Inference 9](#_Toc190278393)

[Objectives 9](#_Toc190278394)

[Hypotheses and Tests 9](#_Toc190278395)

[I. Uniform Distribution of Sailing Transactions: 9](#_Toc190278396)

[II. Association Between Origin and Destination in Discharge Data: 9](#_Toc190278397)

[III. Difference in Average Transaction Month (Discharge vs. Sailing): 10](#_Toc190278398)

[5. Visualization and Presentation of Findings 11](#_Toc190278399)

[Objectives 11](#_Toc190278400)

[Steps Performed 11](#_Toc190278401)

[A. Cumulative Container Analysis (Main Goal): 11](#_Toc190278402)

[B. Distribution and Comparison Visualizations (Other): 12](#_Toc190278403)

[6. Conclusions and Future Directions 14](#_Toc190278404)

[Key Findings 14](#_Toc190278405)

[ Data Structure & Cleaning: 14](#_Toc190278406)

[ Trends in Container Movements: 14](#_Toc190278407)

[ Container Stock Analysis: 14](#_Toc190278408)

[Future Directions 14](#_Toc190278409)

[ Additional Analysis: 14](#_Toc190278410)

[ Refinement: 14](#_Toc190278411)

[ Advanced Techniques: 14](#_Toc190278412)

[ Others: 14](#_Toc190278413)

[THANK YOU! 14](#_Toc190278414)

# Comprehensive Report on Marine Data Analysis

This report documents the analysis process performed on a marine container dataset obtained via an API. The Python code carries out several tasks ranging from data selection, cleaning, exploratory analysis, and statistical testing to final visualizations. The overall aim is to derive actionable insights regarding container movements (imports/exports), seasonal trends, and location-based correlations.

Dataset Citation:  
The data is retrieved from the company’s API

<https://smooth-ocean.tech/gmac-operations/seneca_jcp_mss/api/>

with the proper authentication.

Additional context and documentation for the analysis can be found in the below referenced GitHub Notebook:

[Project1\_ML\_JonathanChacko.ipynb](https://github.com/jcp-tech/Seneca_Class_Notes/blob/master/Semester%201/AIG100%20-%20Machine%20Learing/Project%201/Project%201%20-%20AIG100%20-%20Jonathan%20Chacko.ipynb) (NOTE: This will be Updated by me Regularly when I get Idea’s for Improvement &or when I learn new Concept’s which I can Apply here.)

# 1. Dataset Selection and Preliminary Research

## Goals and Questions

* **Overall Question:**

What insights about container operations (imports/exports) can be derived from the marine dataset?

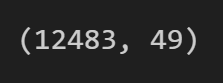
* **Key Questions:**
  + What is the structure of the dataset?
  + How many empty/error values exist, and what cleaning is needed for the analysis?
  + How can we fill missing date values without compromising the data’s integrity or order?
  + How to clean and complete critical port details by leveraging directly correlated columns?
  + What are the correlations between origin and destination in container discharges and sailings?
  + Are sailing transactions uniformly distributed across the 12 months, or do seasonal patterns exist?
  + Is there a statistically significant difference in the average transaction month between discharge and sailing operations?
  + How does the container stock change over time, and in which months is the stock at its highest or lowest?
  + What is the monthly distribution of discharge and sailing transactions?
  + *Other Understanding’s from the Data.*

## Steps Performed

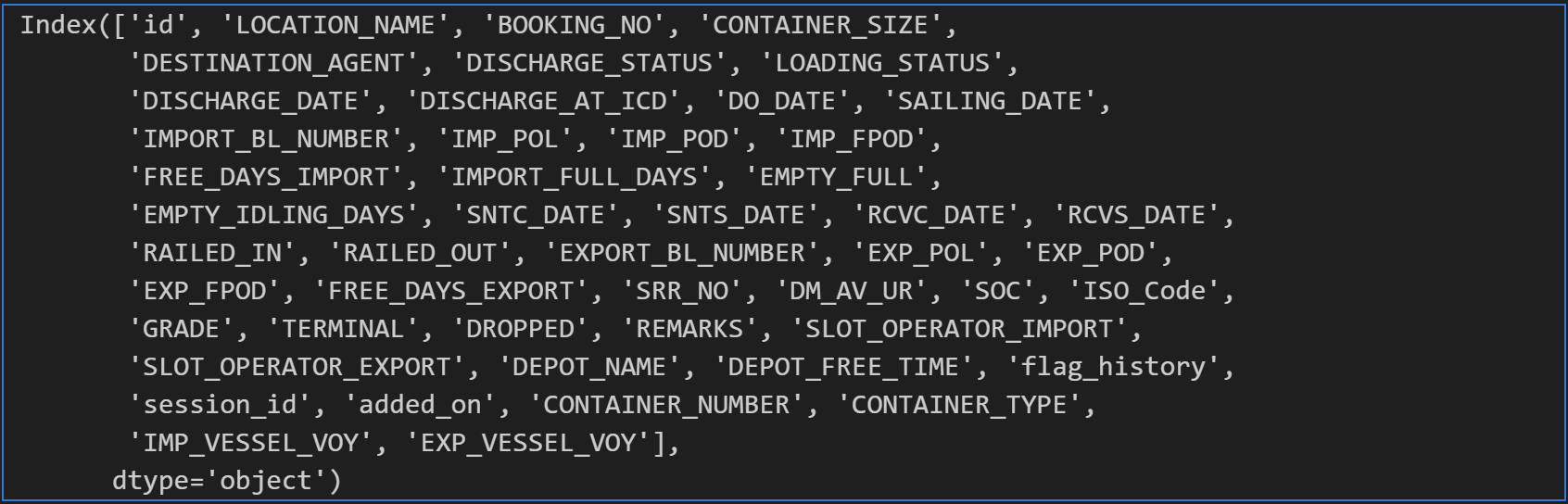
* Dataset Selection:  
  The dataset is selected from a company-controlled API, ensuring that it includes structured data on container movements. A filter is applied to focus on data for specific countries (e.g., SAUDI ARABIA, optionally KUWAIT).
* Preliminary Research:  
  Initial research involves understanding the dataset’s context (marine container operations) and framing the analysis goals as listed above.

## Important Outputs:

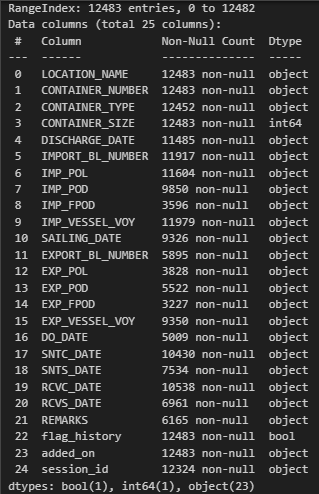
* **Shape:** (Rows, Columns)



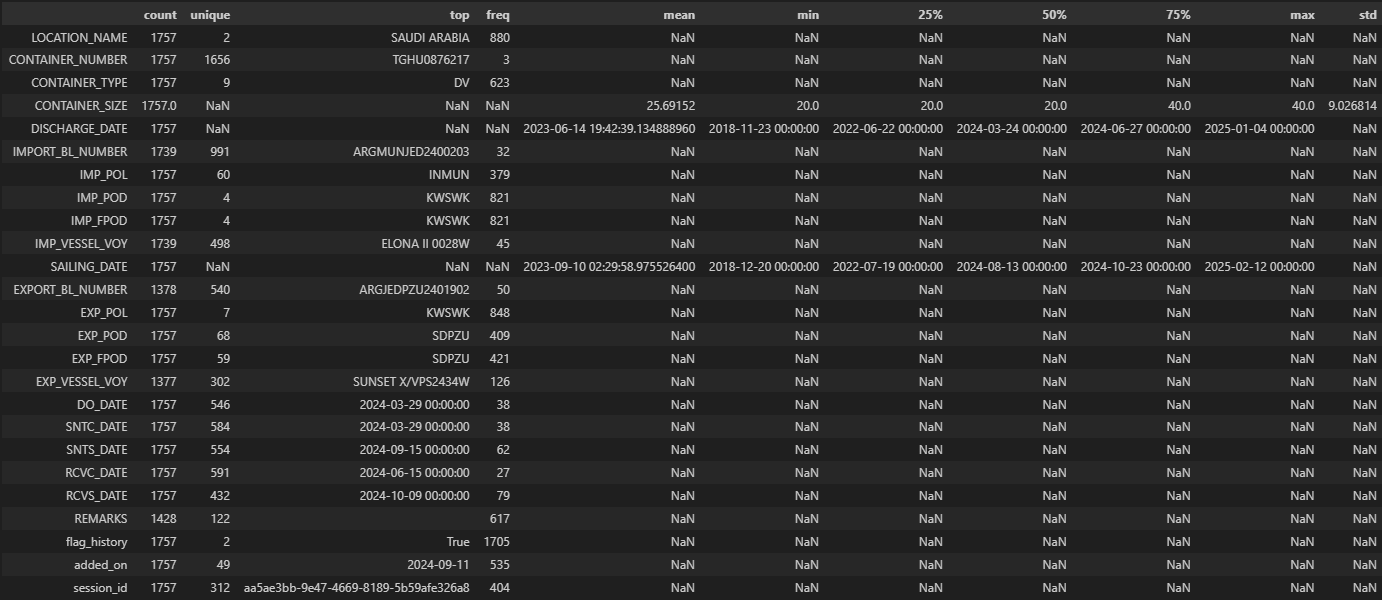
* **Columns** (Excluding some which were Dropped because of Privacy Reason’s)



* **Info** (Including Number of Values not Null & Data Type for each Column)



* **Describe** (Understanding the Data’s which I’ve Transposed Output Given below for Better Visualisation)



# 2. Data Cleaning and Preprocessing

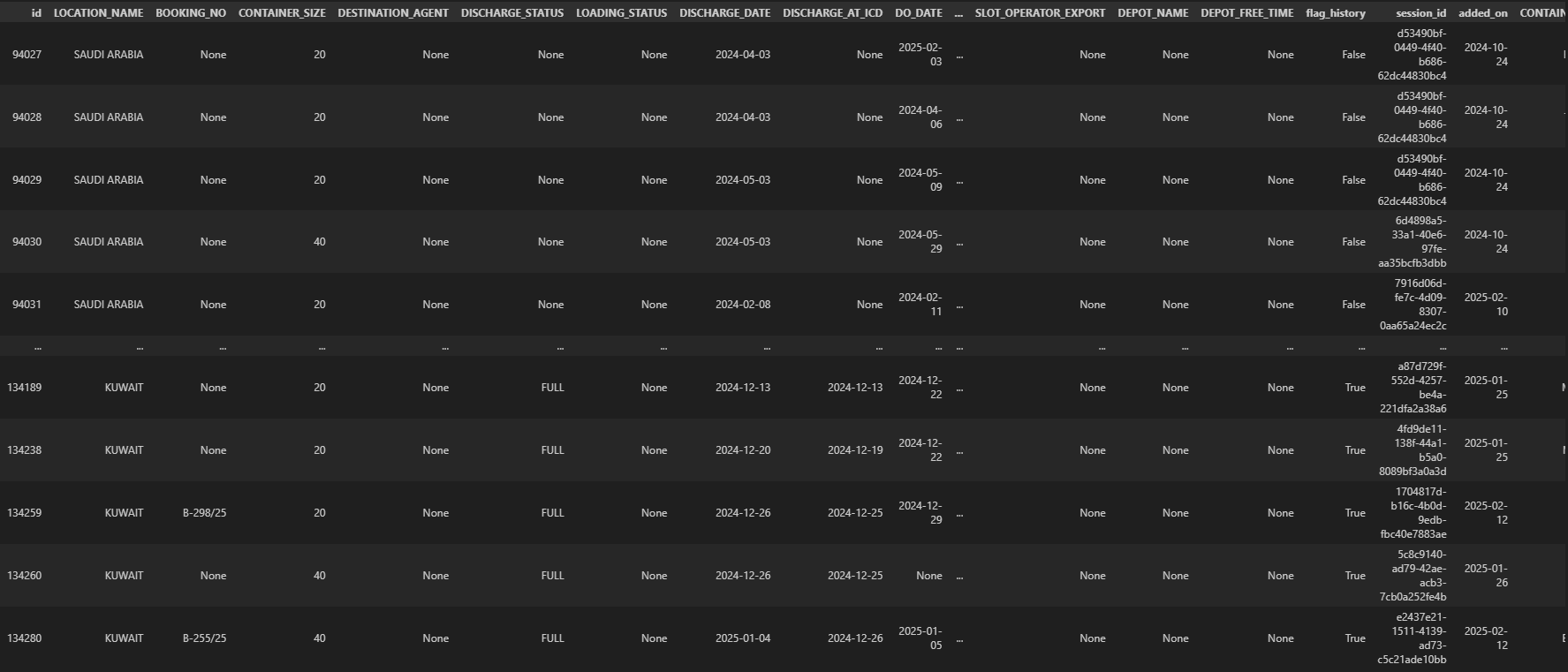
## Objectives

* To prepare the raw data for analysis by handling missing values, erroneous entries, and ensuring correct data types.
* To fill date values and standardize port details for consistency.

## Steps Performed

### Environment Setup and Data Retrieval:

* + Load environment variables (e.g., API key) stored from a `.env` File & called using dotenv Library.
  + Send an API request to retrieve the dataset.
  + Create a DataFrame from the JSON response.



### Initial Data Exploration:

* + Display dataset dimensions, column names, and summary statistics.
  + Identify columns with missing values using df.isnull().sum().

### Date Cleaning and Preprocessing:

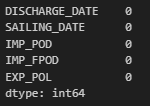
* + **Flag Handling:** Dates with a year before 1691 (flag values, e.g., 1690) are set to None since they indicate placeholder values.
  + **Missing Date Values:**  
    The code fills missing dates using a backward-fill method (bfill), followed by filling any remaining gaps with the current date.  
    **Note:** This ensures that important dates (e.g., DISCHARGE\_DATE and SAILING\_DATE) are not null.
  + **Date Order Validation:**  
    A check is performed to verify that dates are in sequential order from arrival to departure. If not, they are corrected to match the discharge date.

### Data Standardisation:

* + **Port Standardization:**  
    Based on the location (either Kuwait &or Saudi Arabia), the code standardizes port names using defined sets of valid ports.
  + **Filling Missing Values:**  
    For critical port columns (e.g., IMP\_POL, IMP\_POD, EXP\_POL), the script fills missing values using related columns.

### Final Preprocessing Steps:

* + Drop rows with null values in crucial columns (e.g., IMP\_POL, EXP\_POD, EXP\_FPOD) to maintain analysis integrity.
  + Convert key date columns into datetime objects for further analysis.



## 3. Exploratory Data Analysis (EDA)

## Objectives

* To explore the cleaned dataset and identify trends, patterns, and relationships in container operations.
* To visualize the monthly distribution and geographic flow of containers.

## Steps Performed

### Data Filtering for Analysis:

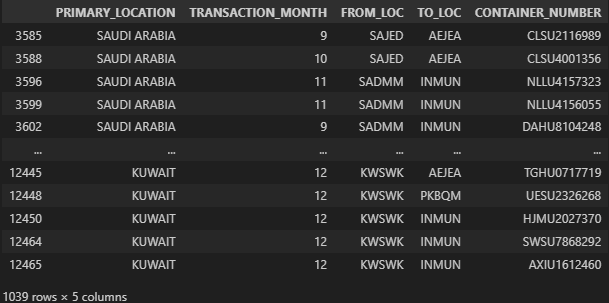
* + Focus on a selected year (e.g., 2024) to analyze a complete period.
  + Split the dataset into two groups:
    - **Discharge Data (Imports):**  
      Rename columns to standardize variables (e.g., DISCHARGE\_DATE becomes TRANSACTION\_DATE, IMP\_POL becomes FROM\_LOC).
    - **Sailing Data (Exports):**  
      Similarly, standardize column names for sailing operations.

### Heatmaps and Pivot Tables:

* + For **Imports:**  
    Group by primary location, transaction month, and origin (FROM\_LOC) to generate a pivot table. Plot a heatmap to visualize the monthly demand.



* + For **Exports:**  
    Group by primary location, transaction month, and destination (TO\_LOC) to visualize export demand via another heatmap.



* + **Outputs:**

|  |  |
| --- | --- |
| **IMPORTS** | **EXPORTS** |
|  |  |

# 4. Statistical Inference

## Objectives

* To statistically test the hypotheses generated from the EDA.
* To determine if there are significant differences or associations within the data.

## Hypotheses and Tests

### Uniform Distribution of Sailing Transactions:

* + **Hypothesis:** Sailing transactions occur uniformly across the 12 months.
  + **Test Used:** Chi-square goodness-of-fit test.
  + **Interpretation:**  
    If the test statistic is high with a low p-value, the null hypothesis (uniform distribution) is rejected.
  + **Interpretation:** (Changes depending on Location Filter)

Chi-square Statistic: 171.29

P-value: 6.40e-31

Although the test statistic is moderate, the p-value is still very small. We reject the null hypothesis, indicating that the data is also significantly non-uniform.

### Association Between Origin and Destination in Discharge Data:

* + **Hypothesis:** There is no association between origin (FROM\_LOC) and destination (TO\_LOC).
  + **Test Used:** Chi-square test of independence.
  + **Interpretation:**  
    A very low p-value would indicate a statistically significant association between the two locations.
  + **Interpretation:** (Changes depending on Location Filter)

Chi-square Statistic: 57.95

P-value: 4.36e-06

Degrees of Freedom: 18

The extremely low p-value indicates a statistically significant association between FROM\_LOC and TO\_LOC. This means that the destination a shipment takes is related to its origin.

Overall, the test provides evidence that shipments (or container discharges) do not move randomly between locations but follow specific, non-independent patterns.

### Difference in Average Transaction Month (Discharge vs. Sailing):

* + **Hypothesis:** There is no difference in the average month of transactions between discharge and sailing operations.
  + **Test Used:** Independent two-sample t-test (Welch’s t-test).
  + **Interpretation:**  
    A statistically significant t-test (small p-value) would indicate that the average transaction month differs between the two groups.
  + **Interpretation:** (Changes depending on Location Filter)

T-test Statistic: -0.11

P-value: 9.09e-01

With a p-value well above the typical 0.05 threshold, there is no statistically significant difference between the mean transaction months for discharge and sailing data.

# 5. Visualization and Presentation of Findings

## Objectives

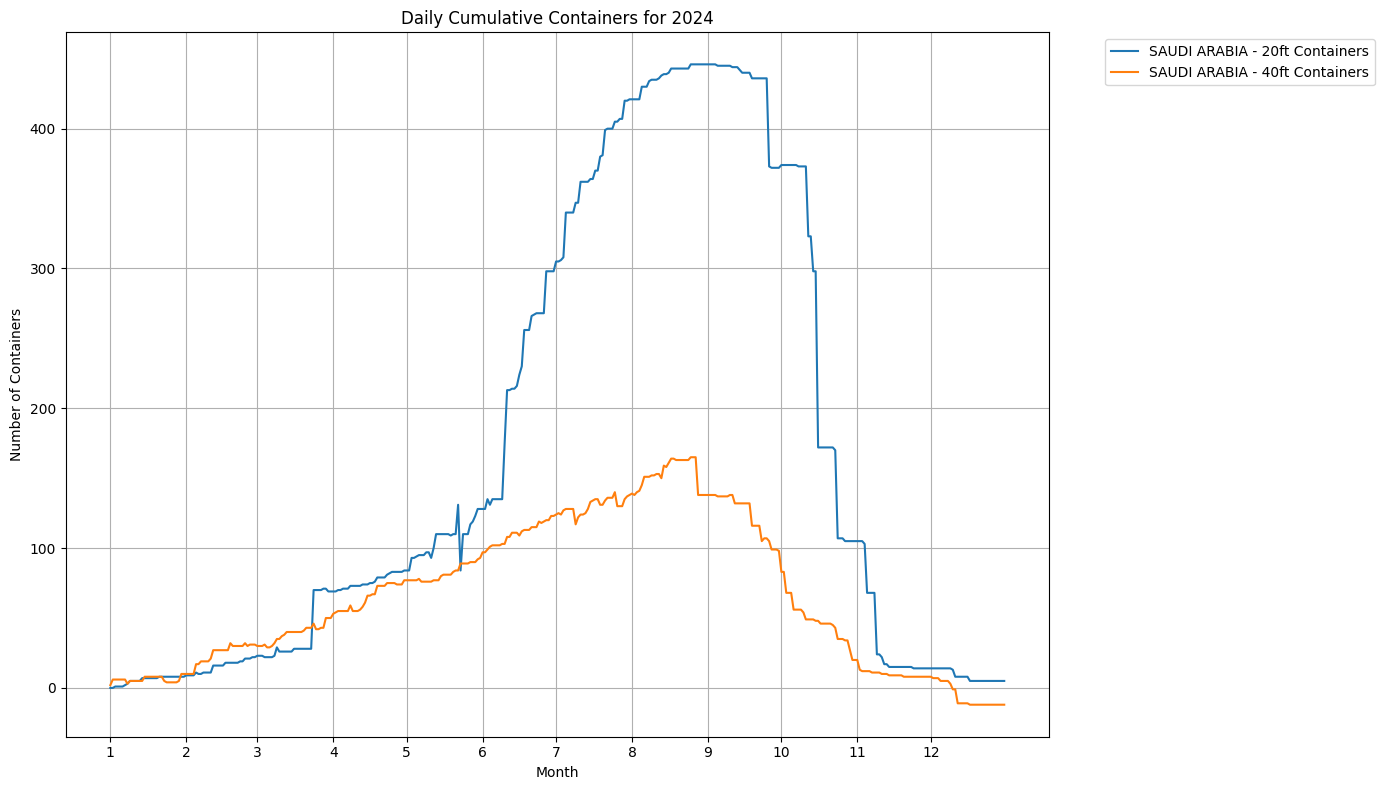
* To clearly visualize the insights and statistical findings.
* To present the cumulative behavior of container operations over time.

## Steps Performed

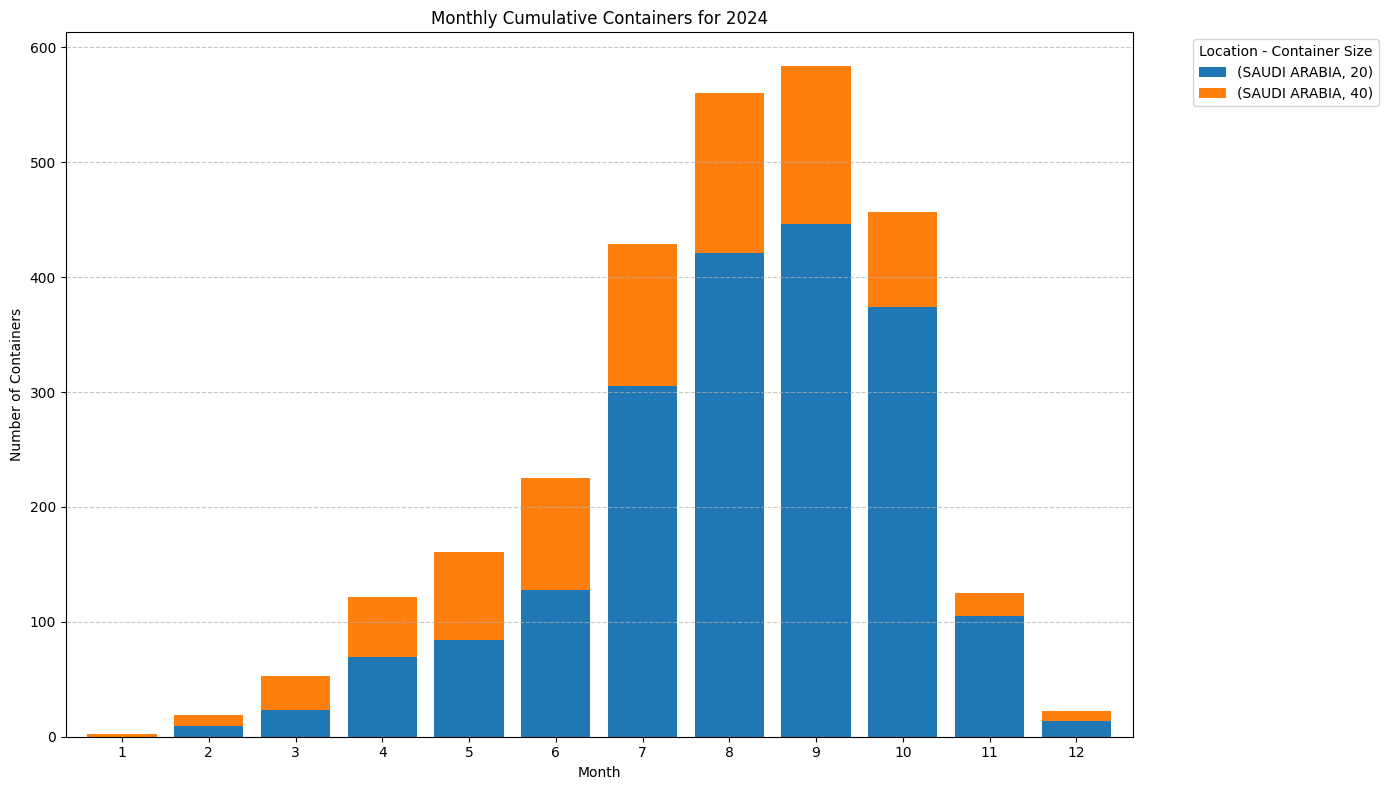
### Cumulative Container Analysis (Main Goal):

* + Compute daily cumulative counts of containers by combining discharge (adding +1) and sailing events (subtracting -1).
  + Group the data by location and container size, then plot:
    - **Time Series Plot:** Daily cumulative container counts.
    - **Monthly Stacked Bar Chart:** Aggregated container counts for the first day of each month.
  + The Goal of this Visualisation is to Understand the Stock Increase/Decrease per Month.

Time Series Plot

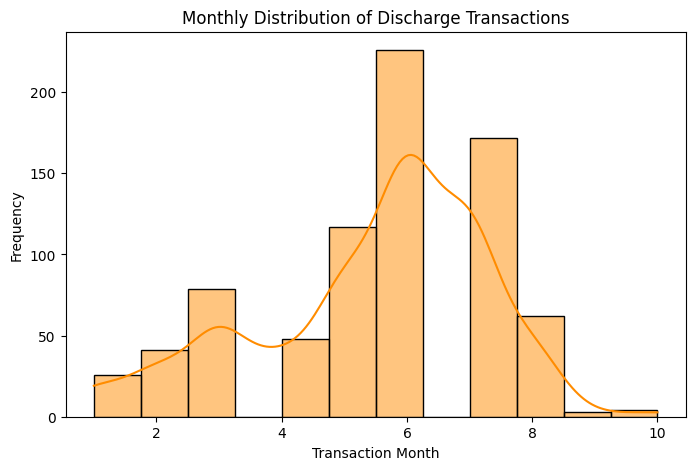
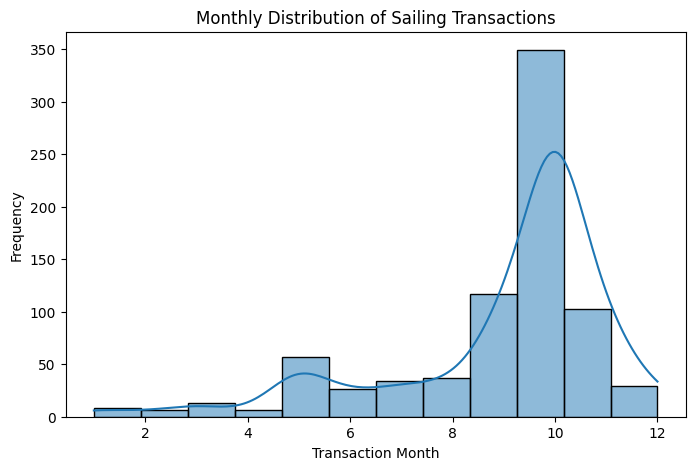
****

Monthly Stacked Bar Chart

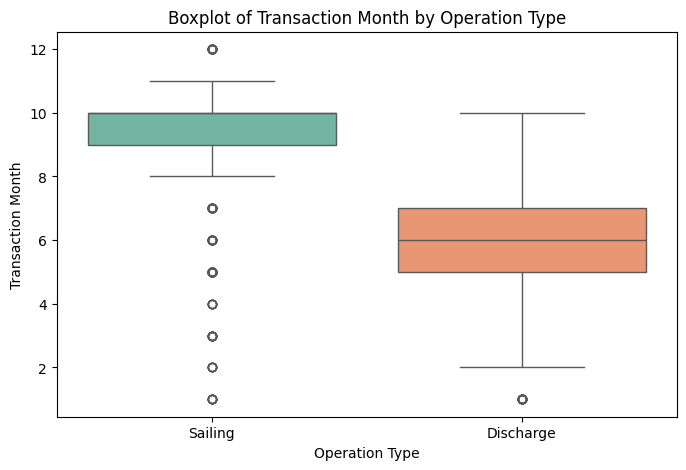


### Distribution and Comparison Visualizations (Other):

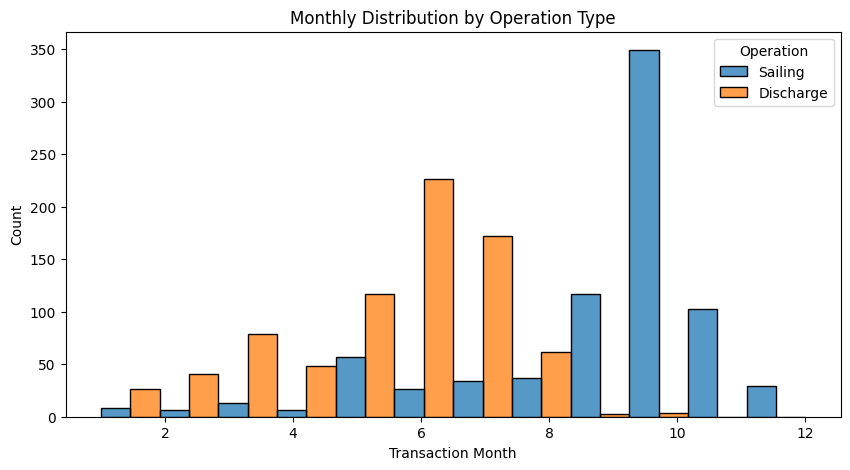
* + **Monthly Distribution Histograms:**  
    Separate histograms for sailing and discharge transactions, enhanced with Kernel Density Estimation (KDE).



* + **Boxplots:**  
    Compare the distribution of transaction months between the two operation types.



* + **Side-by-Side Histograms:**  
    Provide a comparative view of the monthly transaction distribution.



# 6. Conclusions and Future Directions

## Key Findings

* Data Structure & Cleaning:  
  The dataset required significant cleaning—especially around date values and port details—to ensure a reliable analysis.
* Trends in Container Movements:  
  The heatmaps and distribution plots revealed seasonal trends and potential location-based preferences.
* Container Stock Analysis:  
  The cumulative plots offer insight into container stock fluctuations over the year, highlighting periods of low and high inventory.

## Future Directions

* Additional Analysis:  
  Explore deeper relationships (e.g., correlation matrices, time-series forecasting) and incorporate other variables like container types and sizes.
* Refinement:  
  Further refine port standardization and consider additional external datasets to compare against global container trends.
* Advanced Techniques:  
  Consider advanced machine learning methods for predictive analytics regarding container flow and inventory management.
* Others:

If I had more time then when I take the Data to Clean there are other Column’s to clean which I would have Filled Up by Taking Data Available on the Web.

*There is a Site called* [*https://www.bic-code.org/check-digit-calculator/*](https://www.bic-code.org/check-digit-calculator/) *which I am Developing a web Scraper & will Utilise to Fill Up the Null/NaN Values in Columns CONTAINER\_TYPE & CONTAINER\_SIZE.*

# THANK YOU!