**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 proper authentication. Additional context and documentation for the analysis can be found in the referenced GitHub reports:

* [Project Report](https://github.com/jcp-tech/Seneca_Class_Notes/blob/master/Semester%201/AIG100%20-%20Machine%20Learing/Project%201/Report.docx)
* [Reflection Document](https://github.com/jcp-tech/Seneca_Class_Notes/blob/master/Semester%201/AIG100%20-%20Machine%20Learing/Project%201/Reflection.docx)

**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?

**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.

**Output Placeholder:** Insert summary information on the dataset structure (e.g., number of rows/columns, sample records).

**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**

1. **Environment Setup and Data Retrieval:**
   * Load environment variables (e.g., API key) using dotenv.
   * Send an API request to retrieve the dataset.
   * Create a DataFrame from the JSON response.
   * **Placeholder:** *[Insert output of initial DataFrame preview (e.g., df.head() and df.shape) here]*
2. **Initial Data Exploration:**
   * Display dataset dimensions, column names, and summary statistics.
   * Identify columns with missing values using df.isnull().sum().
3. **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.
   * **Placeholder:** *[Insert output showing cleaned date columns and a summary of changes (e.g., before/after null counts)]*
4. **Cleaning Port Details:**
   * **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.
   * **Placeholder:** *[Insert summary table or printout of port columns before and after cleaning]*
5. **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.

**Output Placeholder:** Insert final DataFrame summary (e.g., df.info() and null count summary).

**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**

1. **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.
2. **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.
   * **Placeholder:** *[Insert heatmap outputs for both import and export analysis]*
3. **Additional EDA:**
   * Count and distribution analysis of key variables (e.g., PRINCIPAL\_NAMES distribution).
   * **Placeholder:** *[Insert additional EDA plots, such as histograms or boxplots if applicable]*

**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**

1. **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.
   * **Placeholder:** *[Insert Chi-square test statistic and p-value, along with interpretation]*
2. **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.
   * **Placeholder:** *[Insert contingency table summary, Chi-square statistic, p-value, and interpretation]*
3. **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.
   * **Placeholder:** *[Insert t-test statistic, p-value, and interpretation]*

**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**

1. **Cumulative Container Analysis:**
   * 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.
   * **Placeholder:** *[Insert cumulative container plot outputs]*
2. **Distribution and Comparison Visualizations:**
   * **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.
   * **Placeholder:** *[Insert histogram, KDE plots, and boxplot outputs]*

**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.
* **Statistical Insights:**
  + The chi-square goodness-of-fit test provided evidence on whether sailing transactions follow a uniform distribution.
  + The chi-square test of independence indicated that there is a strong association between origin and destination in the discharge data.
  + The t-test revealed significant differences in the mean transaction months between discharge and sailing operations.
* **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.