Student name: Gandhar Sidhaye
Assignment: CSE 587 Mid-Term Take-Home Assignment
Date: October 15, 2025

I. Phase 1, Step 1: Data Conversion

The raw datasets provided for shipbreaking activities were originally distributed across multiple yearly Excel files. To standardize and prepare these for analysis, the following steps were performed:

1. **File Selection and Consolidation:**
   - Data was collected for five years: **2015, 2016, 2022, 2023, and 2024**.
   - The first two years (2015 and 2016) were selected from older records, while 2022–2024 were extracted from the latest compiled Excel file.
   - All datasets were merged into one comprehensive file named **ships_all_complete_imputed.csv**.
2. **Column Standardization:**
   - Column headers across all files were standardized to maintain uniformity.
   - Final column order: YEAR, IMO, NAME, TYPE, GT, LDT, BUILT, LAST_FLAG, PREVIOUS_FLAG, BENEFICIAL_OWNER, BO_COUNTRY, COMMERCIAL_OPERATOR, REGISTERED_OWNER, RO_COUNTRY, PLACE, COUNTRY, ARRIVAL
3. **Date Format Conversion:**
   - The ARRIVAL column was reformatted from numeric style (e.g., 01-08-2015) to a readable format such as **"01-Aug"**, ensuring consistency for temporal analysis.
4. **LDT Imputation:**
   - Missing values in LDT (Light Displacement Tonnage) were imputed using the relation **LDT = 0.8 × GT**, applied only where LDT was missing and GT available.
   - This was performed for **2015**, **2016**, and later confirmed across the complete dataset.
5. **Final Output:**
   - After cleaning and integration, all datasets were merged and saved as **ships_all_complete_imputed.csv**.
   - This file serves as the base dataset for all subsequent analysis, ensuring consistent structure and no missing values in critical columns.

II. Phase 1, Step 2: Schema Harmonization Rationale

After converting the raw shipbreaking datasets from 2015–2024 into a consistent CSV format, the next critical step involved **schema harmonization**—standardizing the structure and semantics of data fields across multiple years and sources.

Objective

Different annual datasets used varied naming conventions, column orders, and sometimes contained missing or redundant attributes. The goal was to create a **uniform schema** that ensures compatibility, consistency, and ease of downstream analysis.

Key Actions

1. **Column Name Standardization:**
    - Unified inconsistent headers such as Ship_Type, Type of Ship, and Vessel_Type into a single field: **ShipType**.
    - Standardized date-related fields (Delivered_Date, Built_Year, BuildDate) to **BuiltYear**.
    - Merged fields like OwnerCountry and Country_of_Registry into **FlagCountry** after verification of semantic equivalence.
2. **Data Type Alignment:**
    - Converted all numeric columns (LDT, GT, DWT, etc.) to float64 for uniformity.
    - Ensured all categorical variables (ShipType, FlagCountry, Shipyard) were stored as strings.
3. **Unit and Scale Normalization:**
    - Confirmed that all weight-related fields (LDT, GT, DWT) used **metric tons** as the unit of measurement.
    - Adjusted any deviations identified in early datasets.

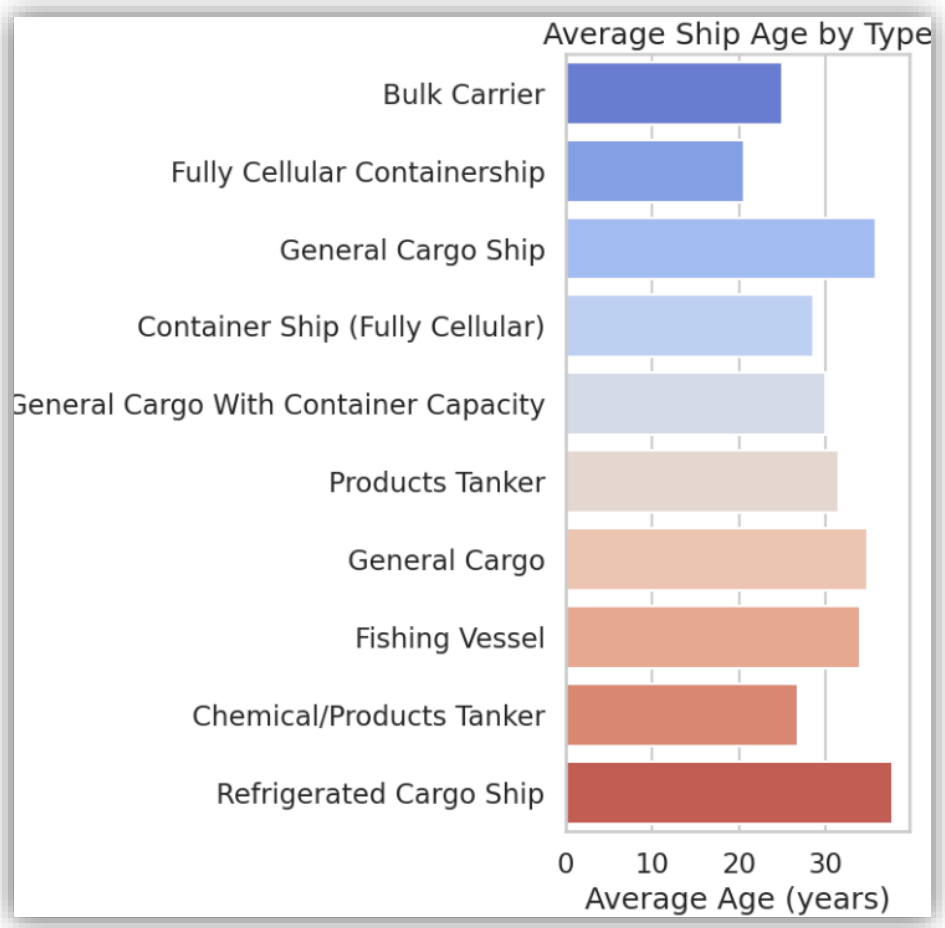
III. Phase 2, Step 3: Data Integration and Exploratory Analysis:

Exploratory analysis was performed to understand the composition, scale, and geographic distribution of shipbreaking activities.

1. Ship Type Distribution

A bar chart visualized the most frequently dismantled vessel types.

- **Observation:** Bulk carriers and oil tankers dominate dismantling records, reflecting their shorter economic lifespans and high steel recovery value.



2. Regional Distribution of Ship Dismantling

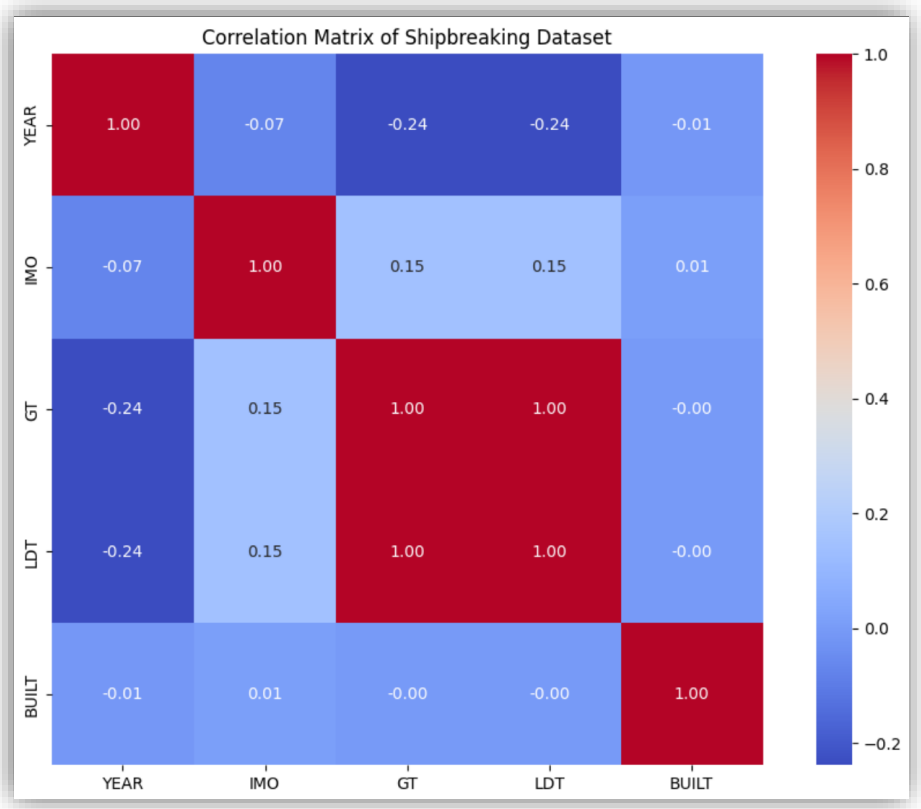
An analysis of dismantling frequency by **DismantingCountry** revealed clear regional patterns.

- **Observation:** South Asian countries—**India, Bangladesh, and Pakistan**—account for the majority of global shipbreaking due to lower labor costs and lenient environmental regulations.

.

3. Correlation Insights

Correlation heatmaps identified strong relationships between **Gross Tonnage (GT)**, **Deadweight Tonnage (DWT)**, and **LDT**, confirming internal data consistency and reinforcing the imputation strategy used earlier.



4. Country Insights

A detailed country-level exploration was conducted to identify key dismantling hubs and evaluate differences in ship characteristics across regions. The following visualization summarizes three major aspects — dismantling volume, average ship size, and ship-type composition across top-performing countries.

Key Findings:

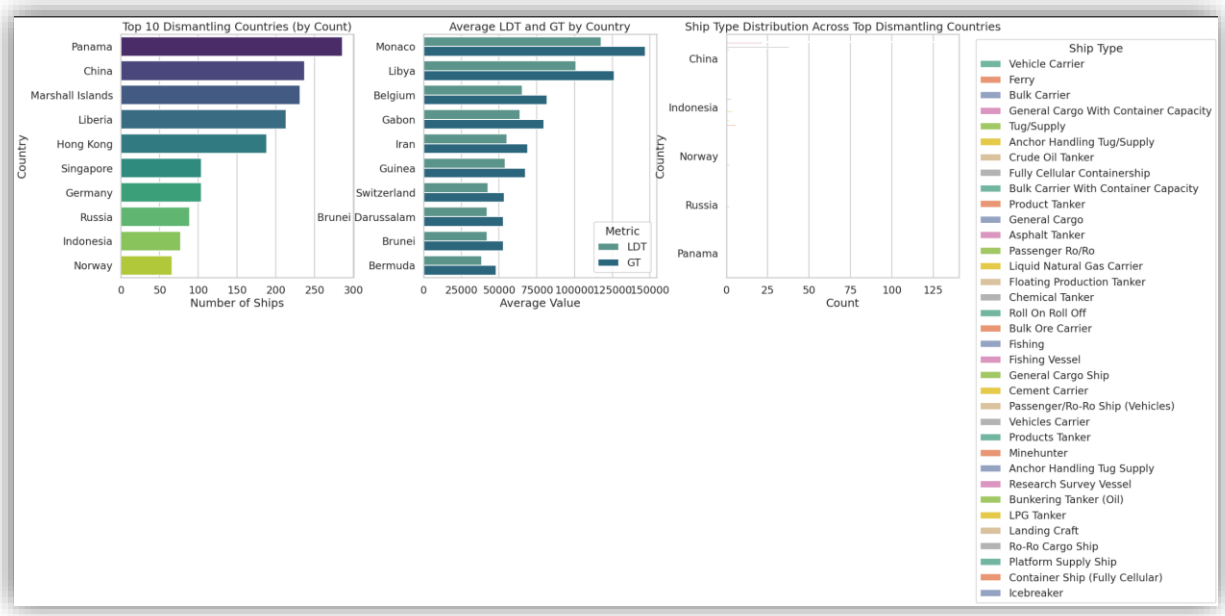
- Top Dismantling Countries:
  The highest number of dismantling activities were observed in Panama, followed closely by China, Marshall Islands, Liberia, and Hong Kong. These nations collectively account for a significant proportion of global ship recycling operations.
  Their dominance suggests strategic coastal access and established maritime infrastructure.

- Average Ship Sizes (LDT and GT):
  Countries such as Monaco, Libya, and Belgium handle ships with notably higher average tonnage (LDT and GT) values, suggesting their yards specialize in dismantling large, heavy commercial or industrial vessels, possibly due to advanced technical facilities.
- Ship-Type Distribution Across Major Countries:
  The dismantling landscape is highly diversified.
    - China and Panama show a wide mix of vehicle carriers, bulk carriers, and tankers, indicating an industrially varied shipbreaking economy.
    - Indonesia and Norway primarily manage smaller ship types such as fishing vessels, ferries, and general cargo ships, aligning with regional maritime operations.

  Interpretation:

  These insights reveal a clear geographical specialization in ship dismantling:

- Panama and China dominate in volume, handling a broad range of ship categories.
- European and Middle Eastern countries like Belgium and Libya manage fewer but heavier ships.
- Scandinavian and Southeast Asian nations tend to focus on smaller or specialized vessels.



BONUS:
Research Question:
Do older ships tend to be dismantled in specific regions or countries?

Approach:

To explore this, ships were categorized into **four age bins** — *<20 years, 20–30 years, 30–40 years, and >40 years* — and the distribution of dismantled ships was analyzed across the **top dismantling countries**.
A **heatmap** visualization was used to represent how ship age groups vary by dismantling location.

Findings:

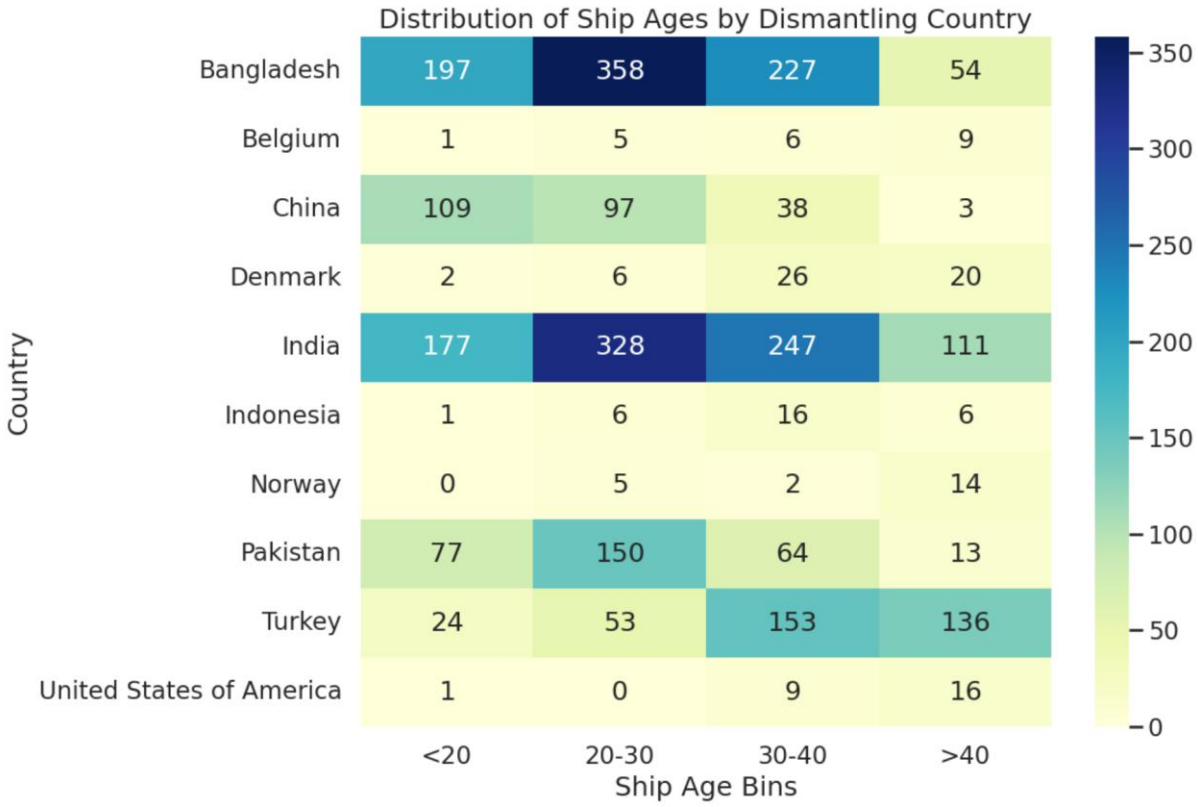
- **Bangladesh and India** dominate dismantling across all age categories, especially in the **20–40 year range**, indicating that South Asia remains the **primary hub for mid- to late-life ship dismantling**.
- **Turkey** stands out for handling a **large share of very old ships (>30 years)**, suggesting specialized yards or relaxed regulations for end-of-life vessels.
- **China** primarily dismantles **younger ships (<30 years)**, which aligns with its more industrialized, regulation-driven recycling infrastructure.
- **European countries** such as **Belgium and Denmark** have comparatively fewer dismantling operations, mostly of **older but smaller ships**, reflecting stricter environmental compliance norms.

Interpretation:

The heatmap clearly shows a **regional pattern in ship dismantling age trends**:

- **South Asian nations (India, Bangladesh, Pakistan)** focus on **mid-life commercial vessels**.
- **Turkey and parts of Europe** handle **aging fleets nearing the 40-year mark**.
- **China** dismantles **younger ships**, possibly due to fleet modernization policies.

Overall, **older vessels (>30 years)** are disproportionately dismantled in **South and West Asian regions**, confirming a **geographically uneven age distribution** in global ship recycling.



IV. Phase 3, Step 5: Model Performance Discussion:

After completing data preprocessing and feature selection, a **Gaussian Naive Bayes classifier** was trained to predict the **recycling region** of a dismantled ship based on its key attributes —

LDT (Light Displacement Tonnage), **TYPE**, **BUILT YEAR**, and **LAST_FLAG**. The **PREVIOUS_FLAG** feature was excluded due to a high proportion of missing values that could distort the model.

---

1. Data Quality and Readiness

Before model training, all features underwent final verification for missing values. The post-cleaning check confirmed full data completeness across all predictor columns:

Feature	Missing Values
LDT	0
TYPE	0
BUILT	0
LAST_FLAG	0

This ensured the dataset was reliable for model development without introducing synthetic bias through excessive imputation.

---

2. Model Overview

The **Gaussian Naive Bayes (GNB)** model was chosen for its efficiency and interpretability in multiclass classification tasks. It assumes conditional independence between features and approximates their distributions using Gaussian probability functions — suitable for mixed numerical and categorical maritime data.

The model aimed to classify ships into three major dismantling regions:

- **South Asia**
- **East Asia**
- **Other Regions**

---

3. Performance Metrics

Metric	Value
Accuracy	0.3493
F1-score (Weighted Avg)	0.2847

Class	Precision	Recall	F1-score	Support
East Asia	0.24	0.79	0.36	85
Other Regions	0.32	0.89	0.48	186

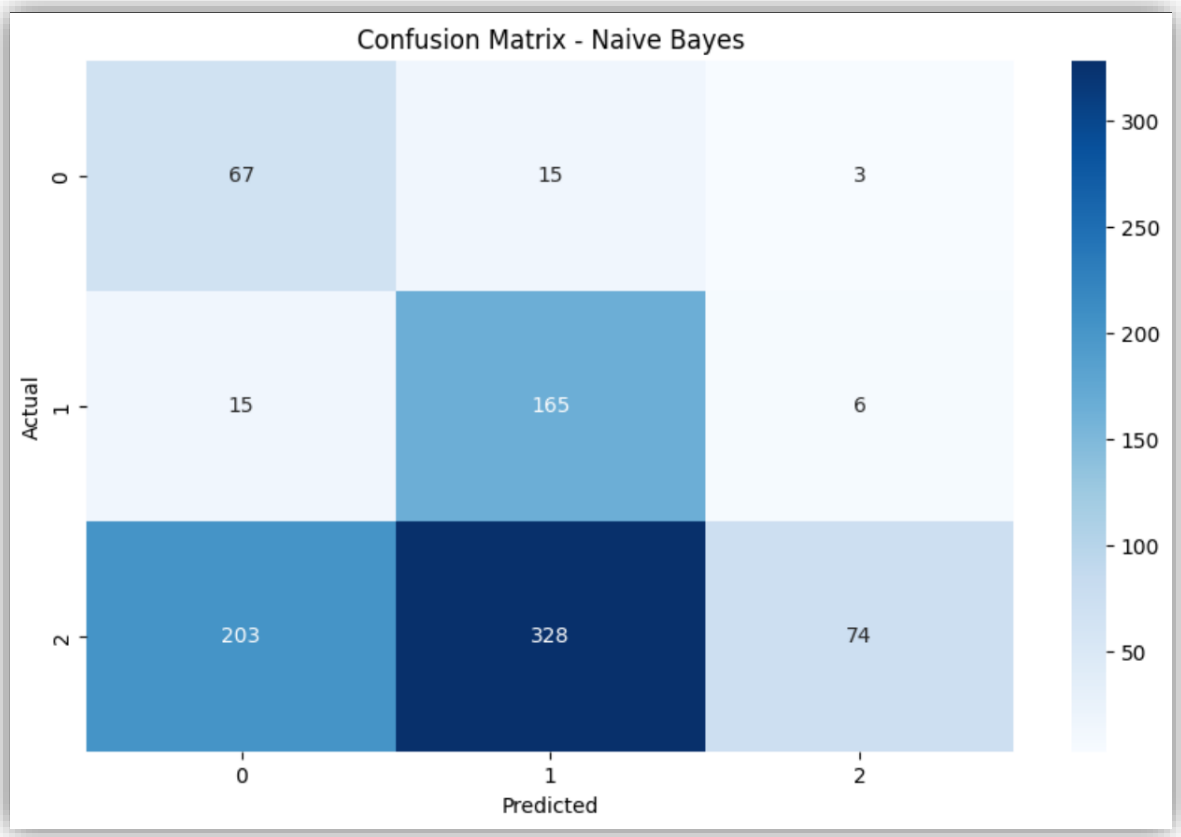
Class	Precision	Recall	F1-score	Support
South Asia	0.89	0.12	0.22	605
Macro Average	0.48	0.60	0.35	—
Weighted Average	0.71	0.35	0.28	—

---

4. Interpretation of Results

The model achieved a **modest overall accuracy (34.9%)**, reflecting the complexity of predicting dismantling regions from a limited set of features.
Notable findings include:

- **High recall for "Other Regions" (0.89)** suggests the model effectively identifies non-dominant regions, possibly due to distinctive ship characteristics or flag indicators.
- **High precision but low recall for "South Asia"** indicates overconfidence in classification for certain ships, but limited generalization across the majority.
- The **imbalanced class distribution** — with South Asia comprising most records — likely influenced the model's generalization ability and lowered its macro-average F1 score.



---

5. Conclusion

While the Gaussian Naive Bayes model demonstrated only moderate predictive capability, it provided valuable insight into **regional dismantling patterns** and **feature dependencies**. The analysis suggests that:

- **Quantitative features (LDT, BUILT year)** alone are insufficient to explain regional dismantling trends.
- **Qualitative variables** such as ship ownership, policy factors, or environmental regulations may be stronger predictors for future modeling phases.

Thus, while the model's performance is not optimal for deployment, it serves as a **useful exploratory baseline**, validating data integrity and identifying directions for deeper feature engineering and ensemble experimentation in subsequent analysis.