SHIP DATASET ANALYSIS

Presented by Dark Knight Developers:
26 - Musab
27 - Anita
28- Anujkumar

Exploratory Data Analysis

Rows: 2736
Columns: 17
DataTypes of column:

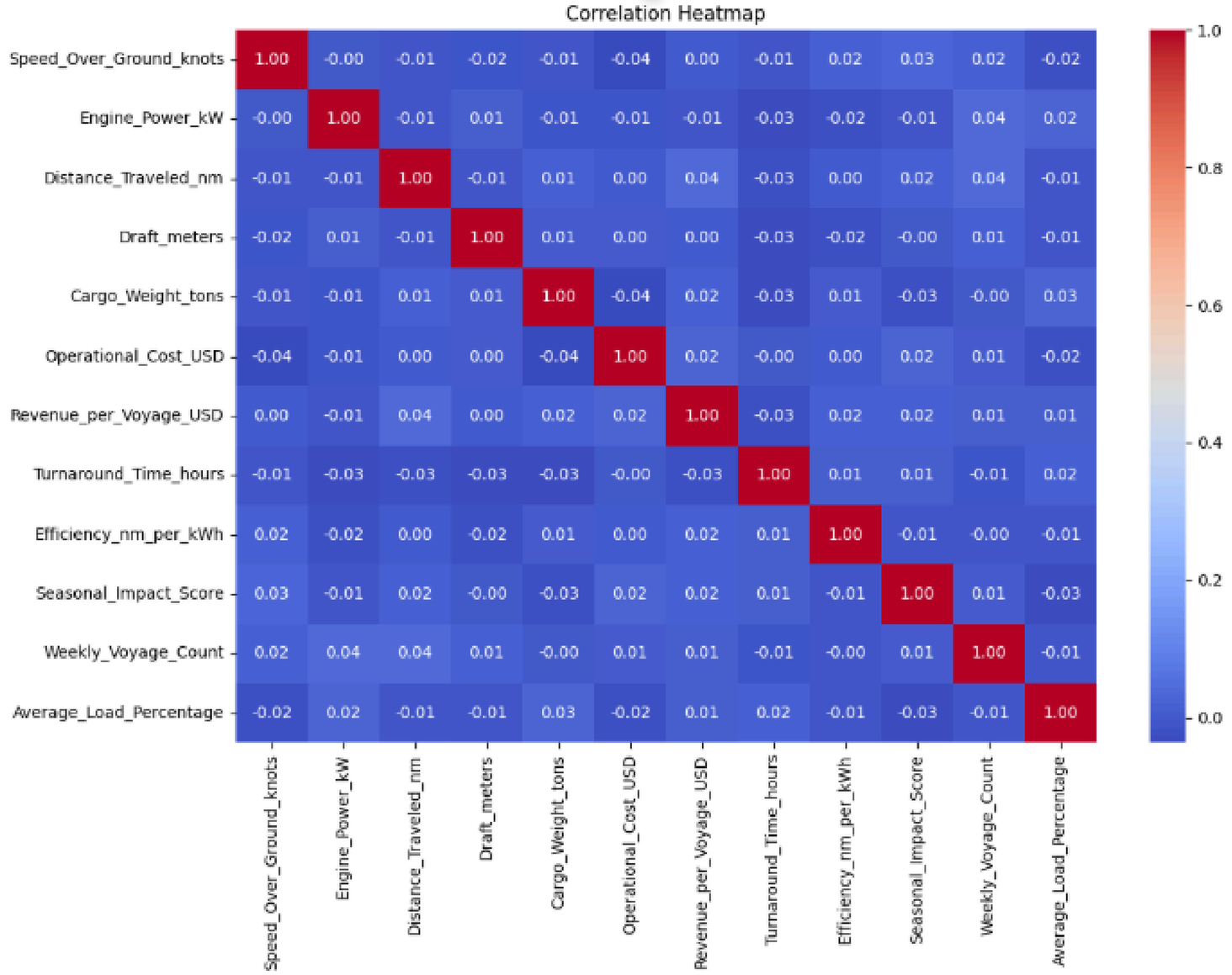
```
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Date                    2736 non-null   object
 1   Ship_Type               2600 non-null   object
 2   Route_Type              2600 non-null   object
 3   Engine_Type             2600 non-null   object
 4   Maintenance_Status      2600 non-null   object
 5   Speed_Over_Ground_knots 2736 non-null   float64
 6   Engine_Power_kW         2736 non-null   float64
 7   Distance_Traveled_nm    2736 non-null   float64
 8   Draft_meters            2736 non-null   float64
 9   Weather_Condition       2600 non-null   object
 10  Cargo_Weight_tons       2736 non-null   float64
 11  Operational_Cost_USD    2736 non-null   float64
 12  Revenue_per_Voyage_USD  2736 non-null   float64
 13  Turnaround_Time_hours   2736 non-null   float64
 14  Efficiency_nm_per_kWh   2736 non-null   float64
 15  Seasonal_Impact_Score   2736 non-null   float64
 16  Weekly_Voyage_Count     2736 non-null   int64
 17  Average_Load_Percentage 2736 non-null   float64
dtypes: float64(11), int64(1), object(6)
memory usage: 384.9+ KB
```

Missing Values:
The dataset had NA values in Weather_Condition, Ship_Type, Route_Type, Engine_Type and Maintenence_Status columns

- **Handled Missing Values using Mode Imputation.**
- **Applied Label Encoding & Min-Max Scaling for consistency.**
- **Extracted Time Features for better analysis.**

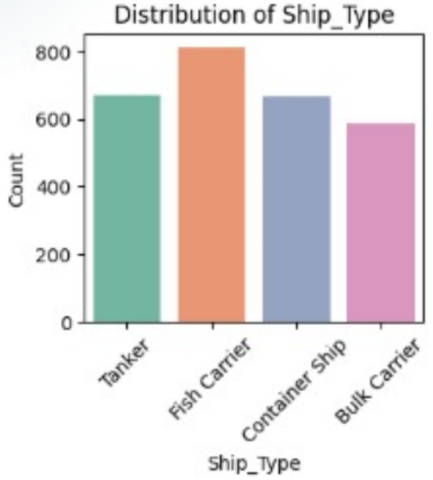
Coorelation Heatmap

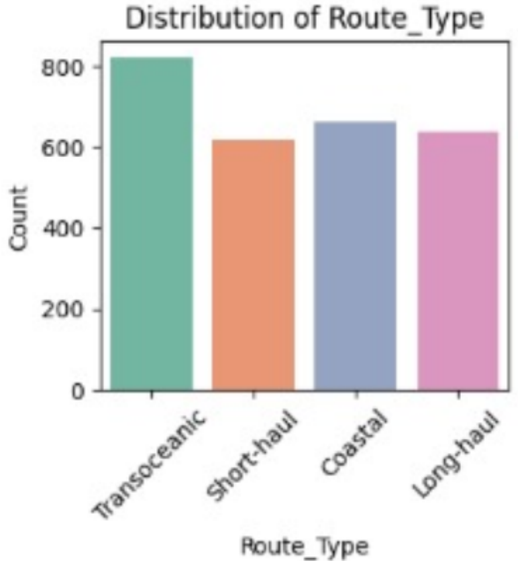
- This heatmap shows the correlation between different variables in our dataset.
- Red (1.0) means strong positive correlation, blue (0.0) means no correlation, and dark blue (-1.0) means strong negative correlation.
- Most correlations are weak, meaning variables do not strongly influence each other.
- The strongest correlation is Draft_meters & Cargo_Weight_tons (1.00), which makes sense as heavier cargo increases draft.
- Thera are no strong relationships
- So using feature selection may help in finding More accurate relationships

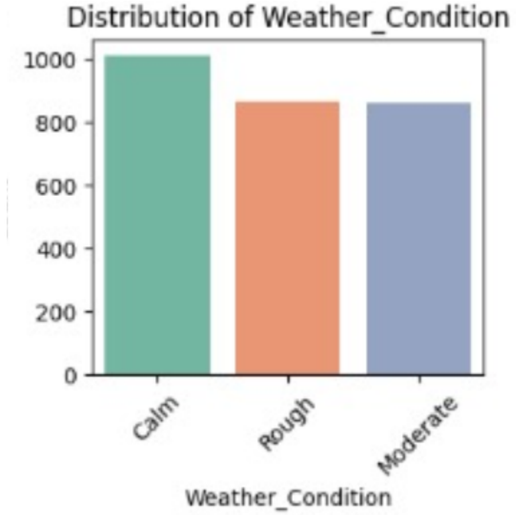


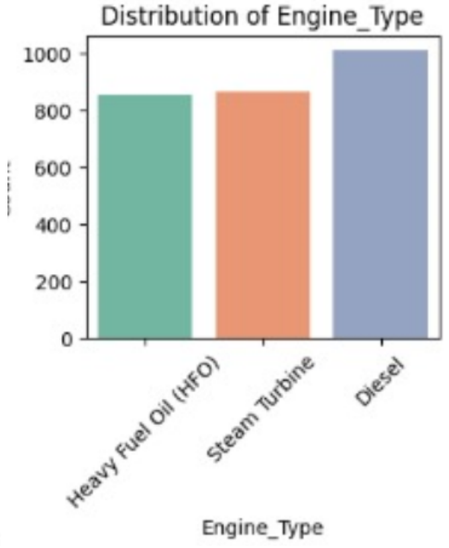
04

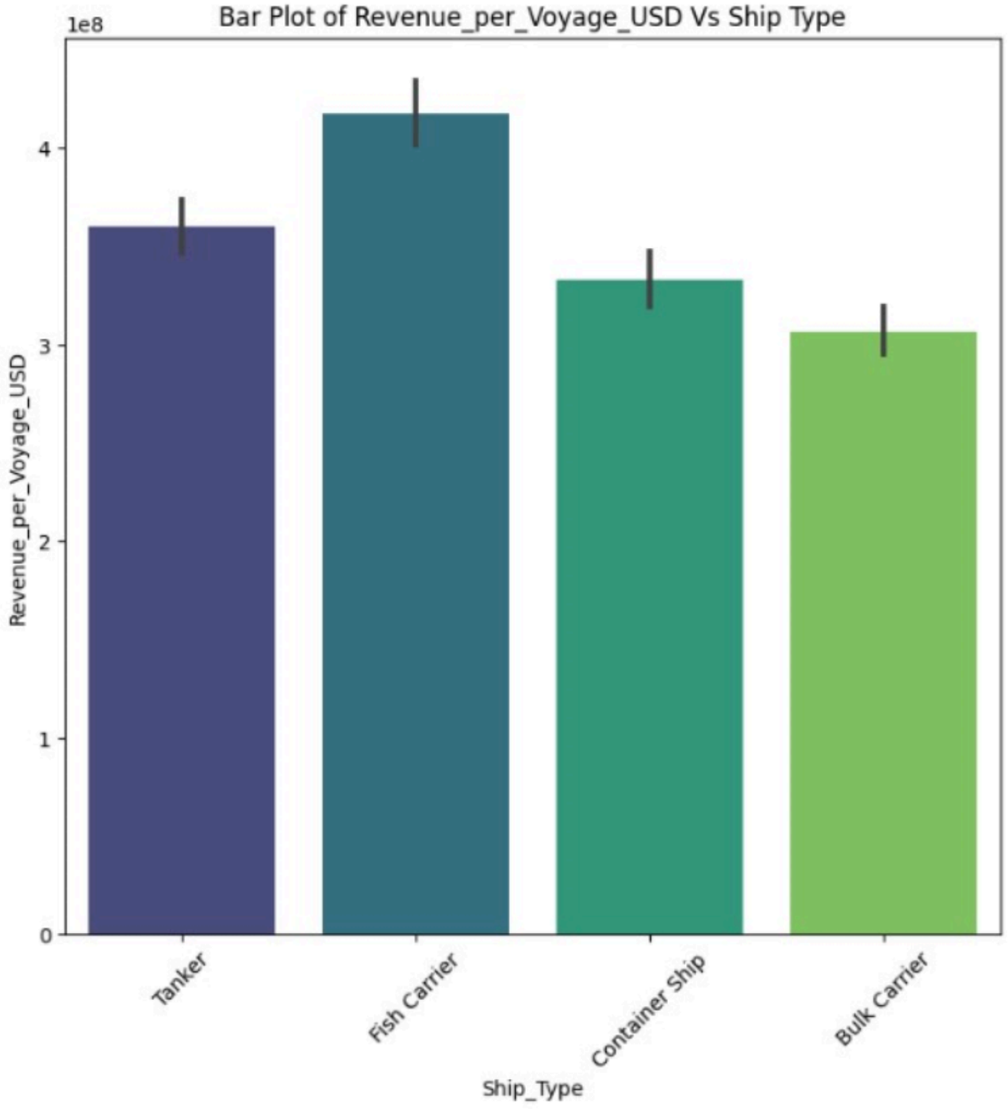
Data Visualizations











Model Building and Evaluation

Problem Statement: Predicting Revenue_Per_Voyage,
Model Used: Multiple Linear Regression
Feature Selection : Lasso Regression(L1)

Model Evaluation:

R-squared: -0.0048
Mean Squared Error: 0.0820
Mean Absolute Error: 0.2497
Root Mean Squared Error: 1.6081

Feature Selected after Lasso(L1):

- Revenue per Voyage in USD
- Engine Type
- Weekly Voyage Count

Lasso:

Revenue_per_Voyage_USD	949564.786887
Engine_Type	0.350886
Weather_Condition	0.122034
Weekly_Voyage_Count	0.022993
Seasonal_Impact_Score	-0.000000
Cargo_Weight_tons	0.000000
Operational_Cost_USD	0.000000
Engine_Power_kW	-0.000000
Turnaround_Time_hours	0.000000
Average_Load_Percentage	-0.000000
Efficiency_nm_per_kWh	-0.000000
Speed_Over_Ground_knots	-0.069903
Draft_meters	-0.076496
Maintenance_Status	-0.332700
Ship_Type	-0.708225
Distance_Traveled_nm	-0.926025
Route_Type	-1.967645

dtype: float64

Model Building and Evaluation

Problem Statement: Operational Cost Prediction
Model Used: Elastic Net
Feature Selection: PCA

Model Evaluation:
Retaining 95% Variance

Mean Squared Error: 0.0846
Mean Absolute Error: 0.2539
R-squared: -0.00015
Root Mean Squared Error: 0.2908

Feature Selected after PCA:

- Ship_Type',
- 'Route_Type',
- 'Engine_Type',
- 'Engine_Power_kW',
- 'Distance_Traveled_nm',
- 'Draft_meters',
- 'Weather_Condition',
- 'Cargo_Weight_tons',
- 'Turnaround_Time_hours',
- 'Average_Load_Percentage

Model Building and Evaluation

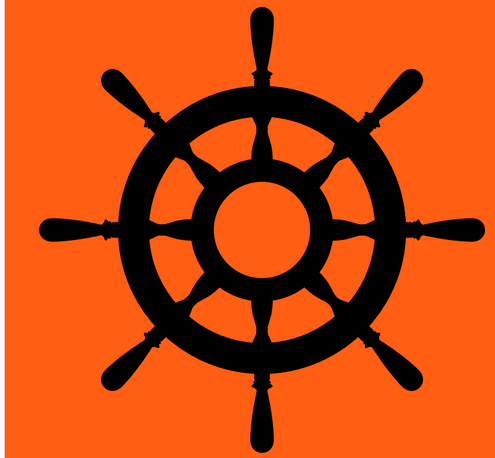
Problem Statement: Predicting Turnaround Time
Model Used:Ridge Regression
Feature Selection: Forward Selection

Model Evaluation:
Alpha(1)

Mean Squared Error: 0.0933
Mean Absolute Error: 0.2703
R-squared: 0.0033
Root Mean Squared Error: 0.3054

Feature Selected after Forward Selection:

- Ship_Type',
- 'Route_Type',
- 'Engine_Type',
- 'Engine_Power_kW',
- 'Distance_Traveled_nm',
- 'Draft_meters',
- 'Weather_Condition',
- 'Cargo_Weight_tons',
- 'Revenue_per_Voyage_USD,
- 'Average_Load_Percentage

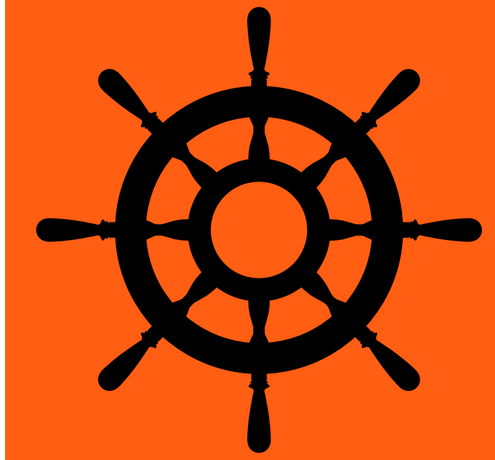


Model Building and Evaluation

Problem Statement: Fuel Effiency Prediction(Efficiency nm per kWh)
Model Used:Linear Regression
Feature Selection: Variance Treshold

Feature Selected

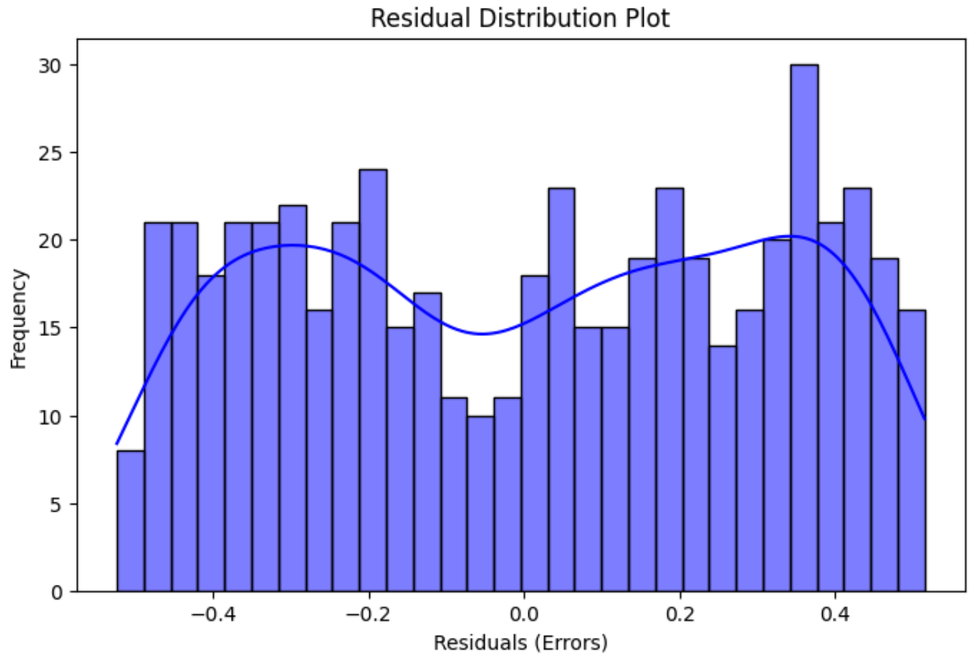
Ship_Type', 'Route_Type', 'Engine_Type', 'Maintenance_Status', 'Speed_Over_Ground_knots', 'Engine_Power_kW','Distance_Traveled_nm', 'Draft_meters', 'Weather_Condition', 'Cargo_Weight_tons', 'Revenue_per_Voyage_USD', 'Turnaround_Time_hours', 'Efficiency_nm_per_kWh', 'Seasonal_Impact_Score', ‘Weekly_Voyage_Count', 'Average_Load_Percentage



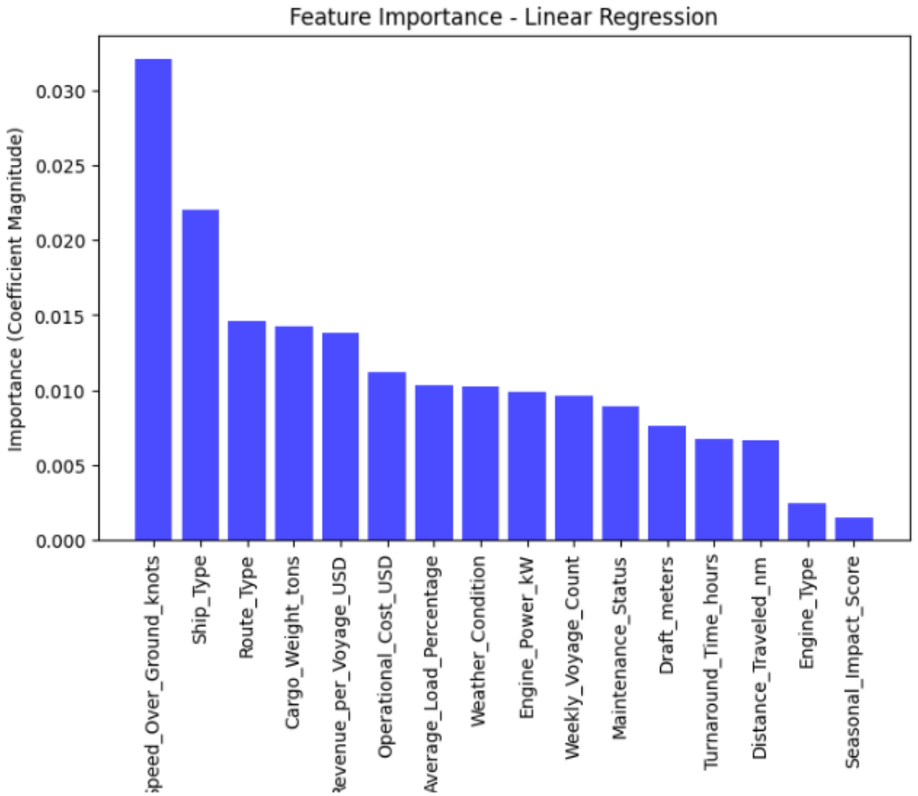
Model Evaluation: Variance Threshold(0.01)

Mean Squared Error: 5.2194395166574736e-31
Mean Absolute Error: 6.424570552156623e-16
R-squared: 1.0
Root Mean Squared Error: 0.2834

Data Visualizations



The residual distribution plot shows that errors are approximately centered around zero, indicating no significant bias in predictions. However, the irregular shape and multiple peaks suggest potential deviations from normality, which may indicate model limitations or unaccounted patterns in the data.



This Feature Importance Bar Chart indicates the relative contribution of each feature in predicting Fuel Efficiency (nm per kWh) using Linear Regression. The most influential feature is Speed_Over_Ground_Knots, followed by Ship_Type and Route_Type, while Seasonal_Impact_Score has the least impact.

Conclusion

- **This analysis explored key maritime metrics like Operational Cost, Turnaround Time, and Fuel Efficiency using regression models. Feature selection and scaling improved predictions, but weak correlations limited accuracy.**
- **The strongest correlation was between Draft and Cargo Weight, while Fuel Efficiency showed perfect accuracy ($R^2 = 1.0$), with Speed Over Ground Knots as the key factor.**
- **Challenges like low R^2 values suggest the need for advanced models and real-time data integration.**

Scan this QR code to see live models deployed on Streamlit:

or

Click Here!