AUTOMATIC DATA RECOGNITION

The research question: Is it possible to predict fuel consumption based on various shipment parameters.

Chen Shmuel

Tal Tubul

INTRODUCTION

In this project, we will use a given data frame containing approximately 1659 rows and 35 features each. The rows of the data represent many shipment parameters as well as general information, such as Vessel, Voyage, Steam, Speed and different oil consumptions. we build a column of total oil consumption. In this project, we will try to predict this total oil consumption based on the features.

STAGES

- Features Investigation
- Data Investigation and Pre-Analsys
- Features Engineering
- EDA and Visualization
- Correlation findings
- Data Preparation
- Machine Learing - Classification models
- Conclusion

FEATURES

- Vessel
- Voyage
- Dated
- Steam
- RPM
- Speed
- STW
- M/E HFO CONS
- M/E LSHFO CONS
- M/E MDO CONS
- M/E LSMDO CONS
- BOIL. LSMDO CONS
- BOIL. MGO CONS
- M/E MGO CONS
- M/E LSMGO CONS
- D/G HFO CONS
- D/G LSHFO CONS
- D/G MDO CONS
- D/G LSMDO CONS
- D/G MGO CONS
- D/G LSMGO CONS
- BOIL. HFO CONS
- BOIL. SHFO CONS
- BOIL. MDO CONS
- BOIL. LSMGO CONS
- WIND
- SWELL
- SLIP
- CUR SPD
- DRAFT FOR
- DRAFT AFT
- OBS DIST
- OPERATION
- CARGOCARRIED

Steam

This column contains the steam hours, which is the number of hours the ship's steam engine was in operation. Measurement - hours.

FEATURES INVESTIGATION

WIND

This column contains information about the wind conditions during the voyage.

Measurement - knots or meters per second.

Normal values: 0 - 50 knots

THE DATA FRAME

	VESSEL	VOYAGE	DATED	TO	STEAM	RPM	SPEED	STW	M/E HFO CONS.	M/E LSHFO CONS.	...	BOIL. LSMGO CONS.	WIND	SWELL	SLIP	CUR SPD	DRAFT FOR	DRAFT AFT	OBS DIST	OPERATION
0	BRAVERUS	201805L	31/12/2018	SINGAPORE PEGBC	24.00	69.0	10.37	11.38	35.3	0.0	...	0.0	5	2.0	17.41	1.0	16.00	16.36	249.0	NaN
1	BRAVERUS	201805L	30/12/2018	SINGAPORE PEGBC	23.00	69.1	9.86	11.07	34.5	0.0	...	0.0	5	2.2	21.56	1.2	16.00	16.36	227.0	NaN
2	BRAVERUS	201805L	29/12/2018	SINGAPORE PEGBC	24.00	70.1	10.20	11.21	35.8	0.0	...	0.0	6	2.0	20.03	1.0	16.00	16.36	245.0	NaN
3	BRAVERUS	201805L	28/12/2018	SINGAPORE PEGBC	23.00	70.0	11.08	11.49	35.3	0.0	...	0.0	5	1.5	13.02	0.4	16.00	16.36	255.0	NaN
4	BRAVERUS	201805L	27/12/2018	SINGAPORE PEGBC	23.83	69.8	10.74	11.49	35.5	0.0	...	0.0	5	1.5	14.29	0.6	16.00	16.36	256.0	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1654	BRAVERUS	201501B	04/01/2015	SINGAPORE	1.00	68.0	10.00	0.00	1.1	0.0	...	0.0	1	0.1	0.00	1.0	6.35	8.27	10.0	NaN
1655	BRAVERUS	201501B	04/01/2015	CAOFEIDIAN	0.00	0.0	0.00	0.00	1.0	0.0	...	0.0	0	0.0	0.00	0.0	6.35	8.27	0.0	DISCHARGING
1656	BRAVERUS	201408L	03/01/2015	CAOFEIDIAN	0.00	0.0	0.00	0.00	0.0	0.0	...	0.0	0	0.0	0.00	0.0	8.48	9.94	0.0	DISCHARGING
1657	BRAVERUS	201408L	02/01/2015	CAOFEIDIAN	0.00	0.0	0.00	0.00	0.0	0.0	...	0.0	0	0.0	0.00	0.0	11.73	12.78	0.0	DISCHARGING
1658	BRAVERUS	201408L	01/01/2015	CAOFEIDIAN	0.00	0.0	0.00	0.00	1.5	0.0	...	0.0	0	0.0	0.00	0.0	17.52	17.67	0.0	DISCHARGING

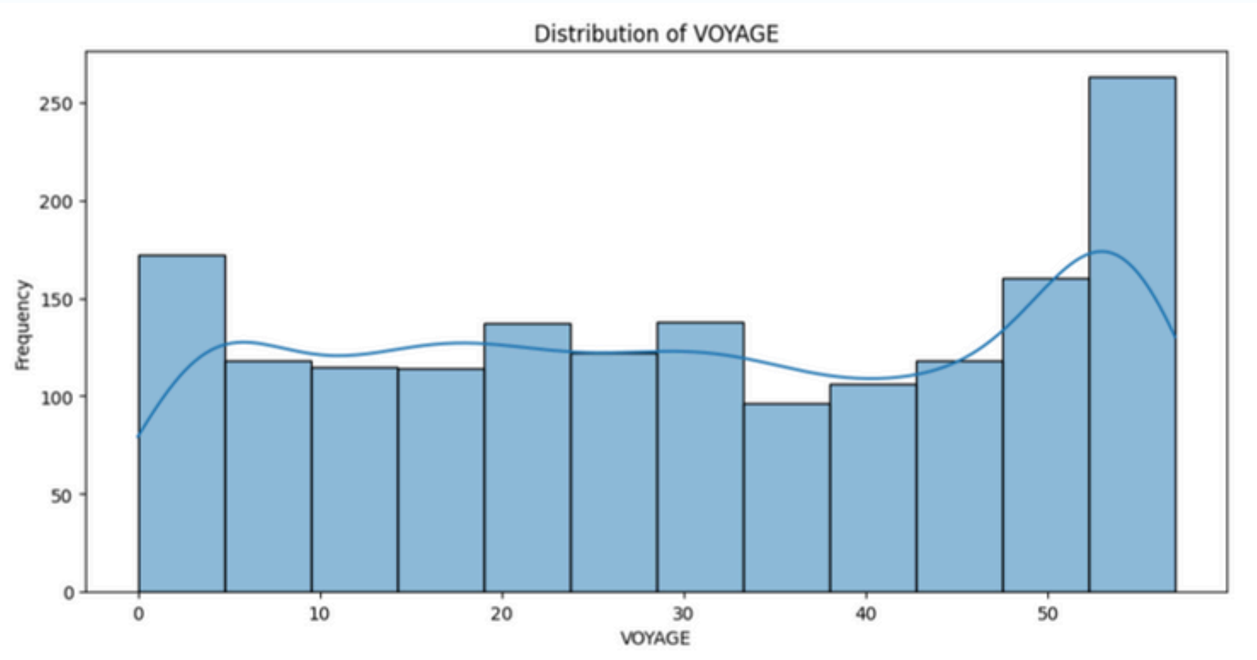
DATA INVESTIGATION

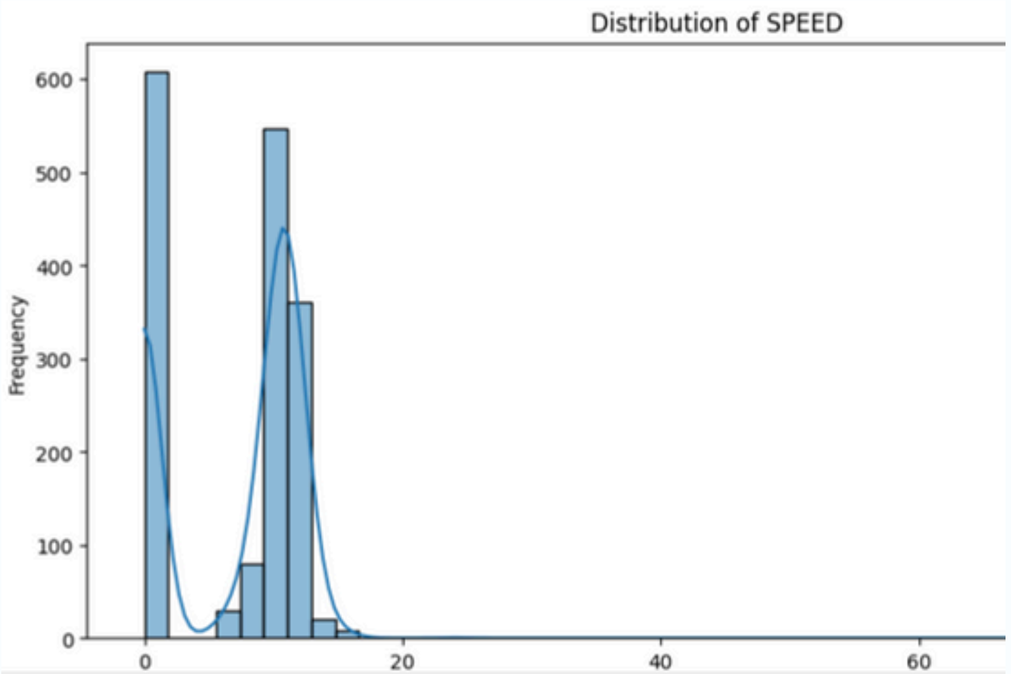
Conclusions

- **17 features**
- 1695 rows
- 28815 raw data
- There is no data < 0
- There is no missing data
- 13 Continuous variables
- 4 Catogrial variables

```
VOYAGE is categorical and contains 58 unique values
DATED is categorical and contains 1459 unique values
TO is categorical and contains 40 unique values
STEAM is continuous (numeric)
RPM is continuous (numeric)
SPEED is continuous (numeric)
STW is continuous (numeric)
WIND is continuous (numeric)
SWELL is continuous (numeric)
SLIP is continuous (numeric)
CUR SPD is continuous (numeric)
DRAFT FOR is continuous (numeric)
DRAFT AFT is continuous (numeric)
OBS DIST is continuous (numeric)
OPERATION is categorical and contains 6 unique values
CARGOCARRIED is continuous (numeric)
summed oil parameters is continuous (numeric)
```

EDA AND VISUALISATION

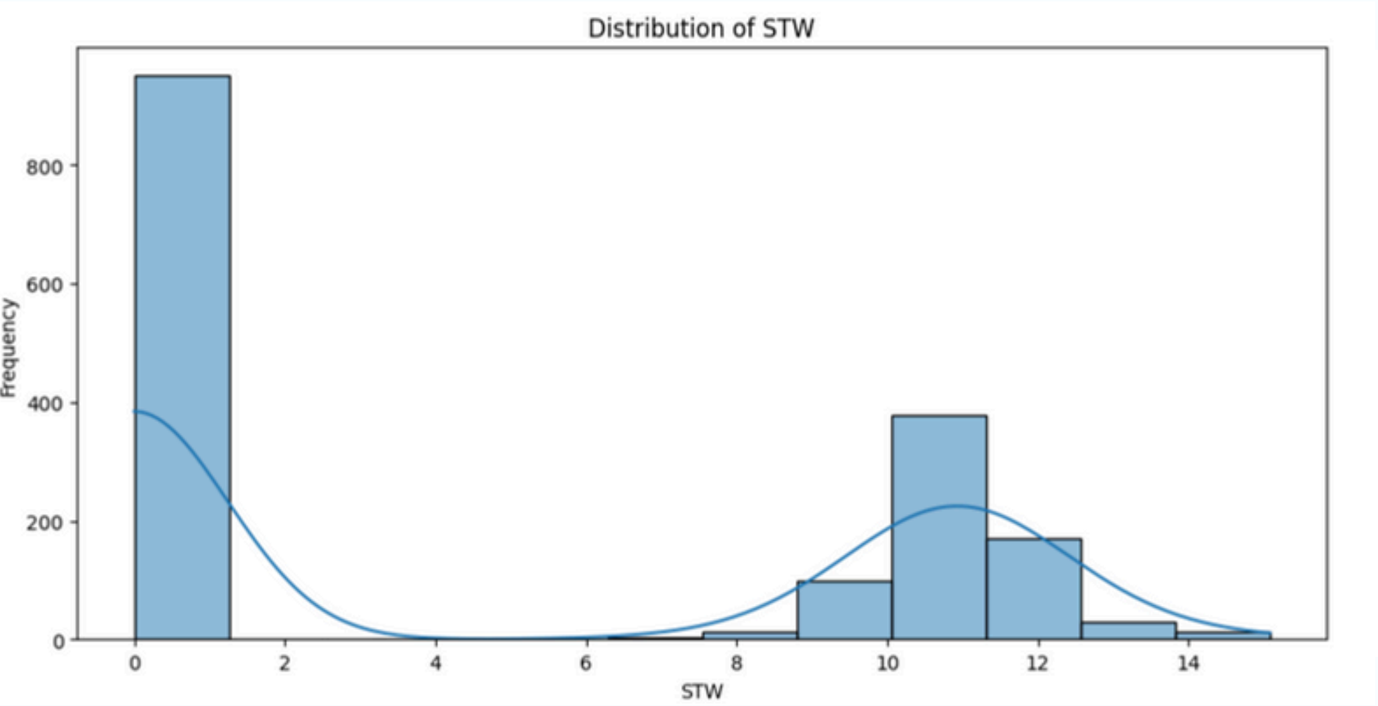


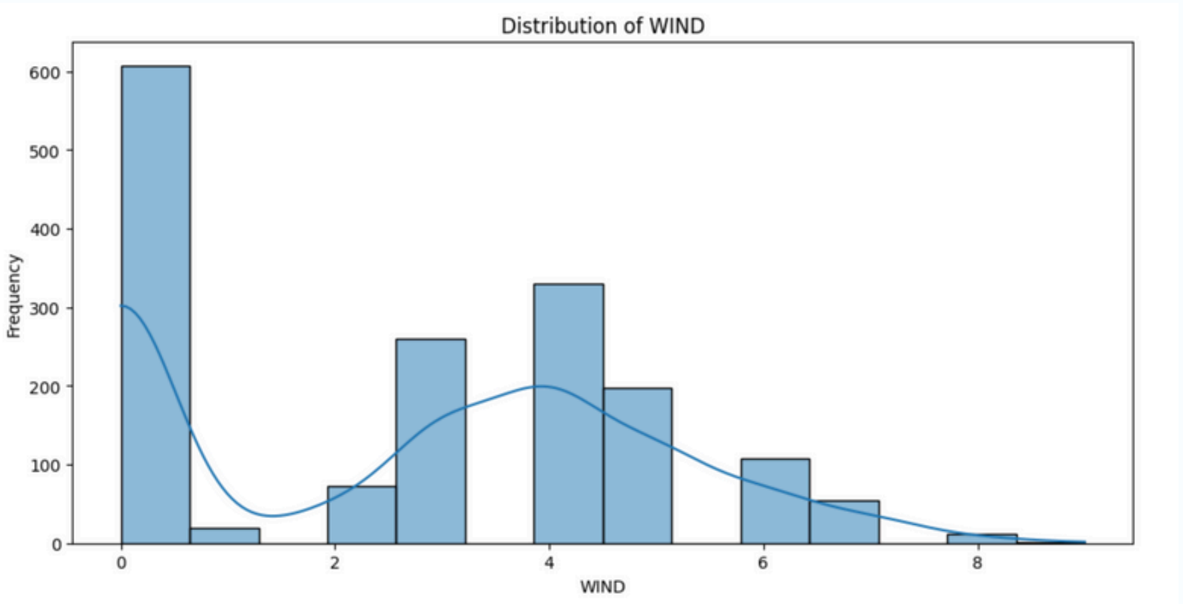


Using the histplot graph
To show the distribution of VOYAGE

Using the histplot graph
To show the distribution of SPEED

EDA AND VISUALISATION

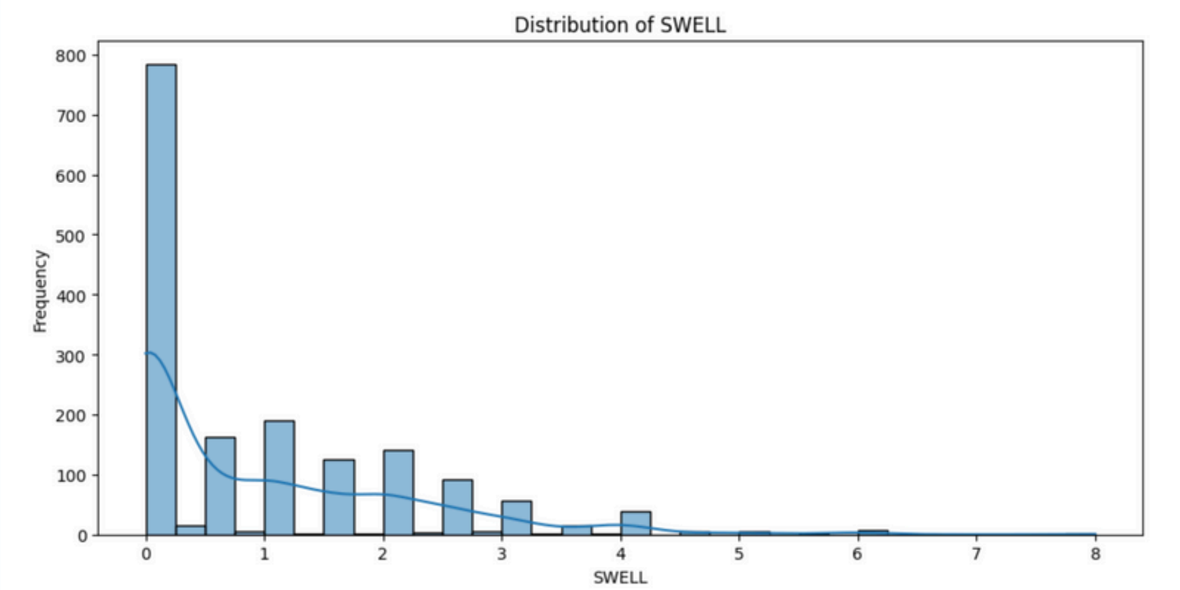


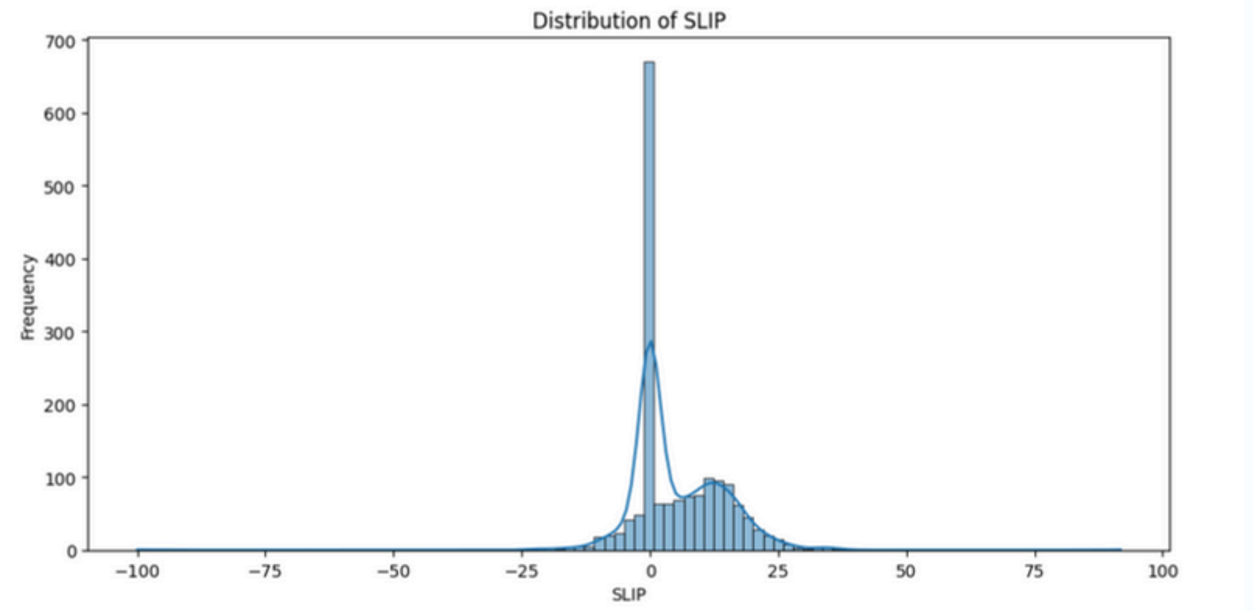


Using the histplot graph
To show the distribution of STW

Using the histplot graph
To show the distribution of WIND

EDA AND VISUALISATION





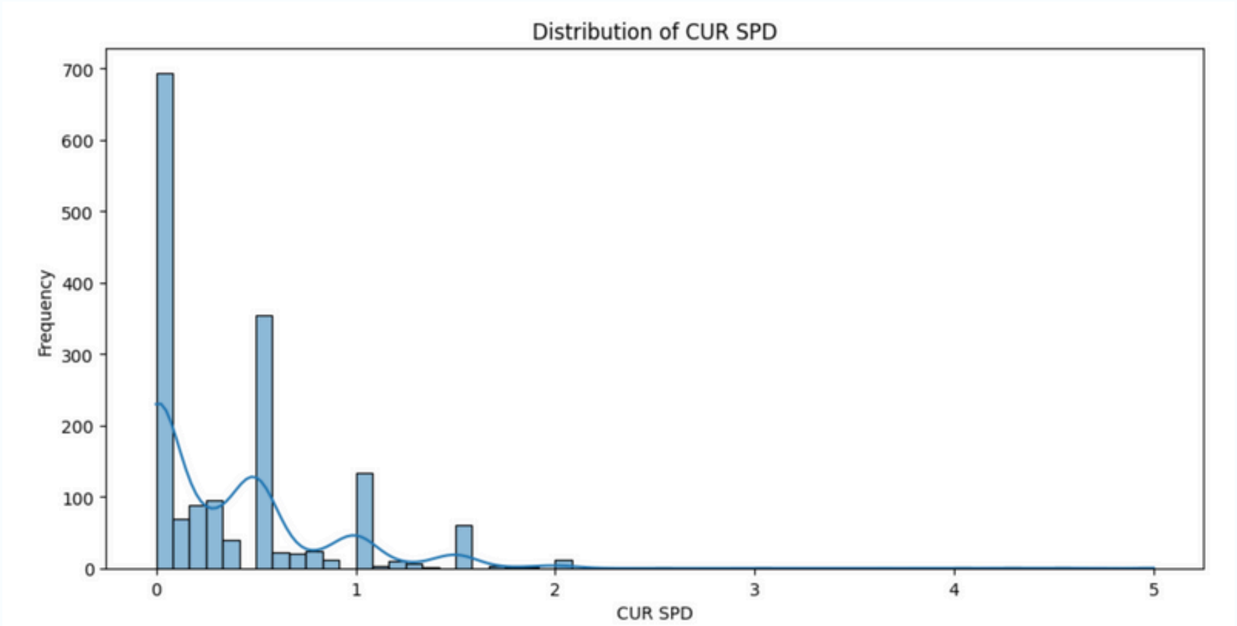
Using the histplot graph

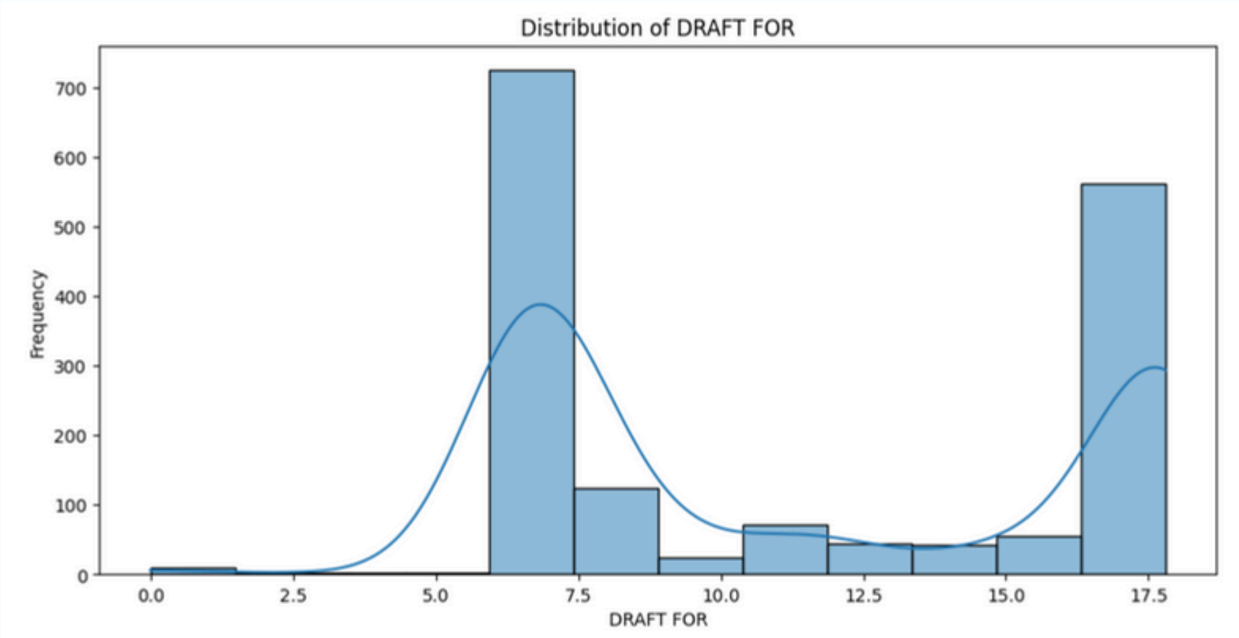
To show the distribution of SWELL

Using the histplot graph

To show the distribution of SLIP

EDA AND VISUALISATION





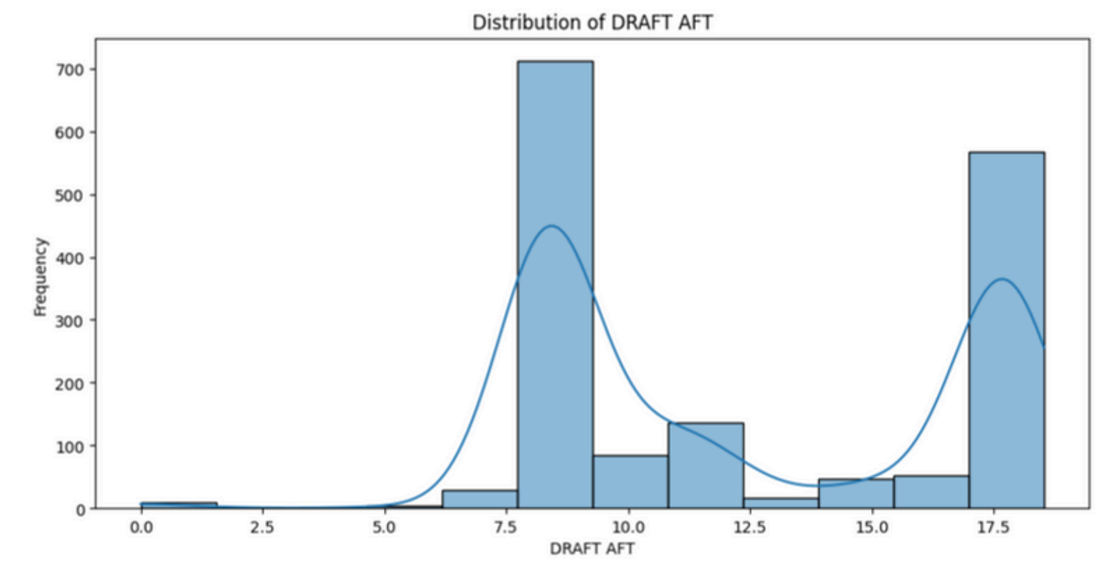
Using the histplot graph

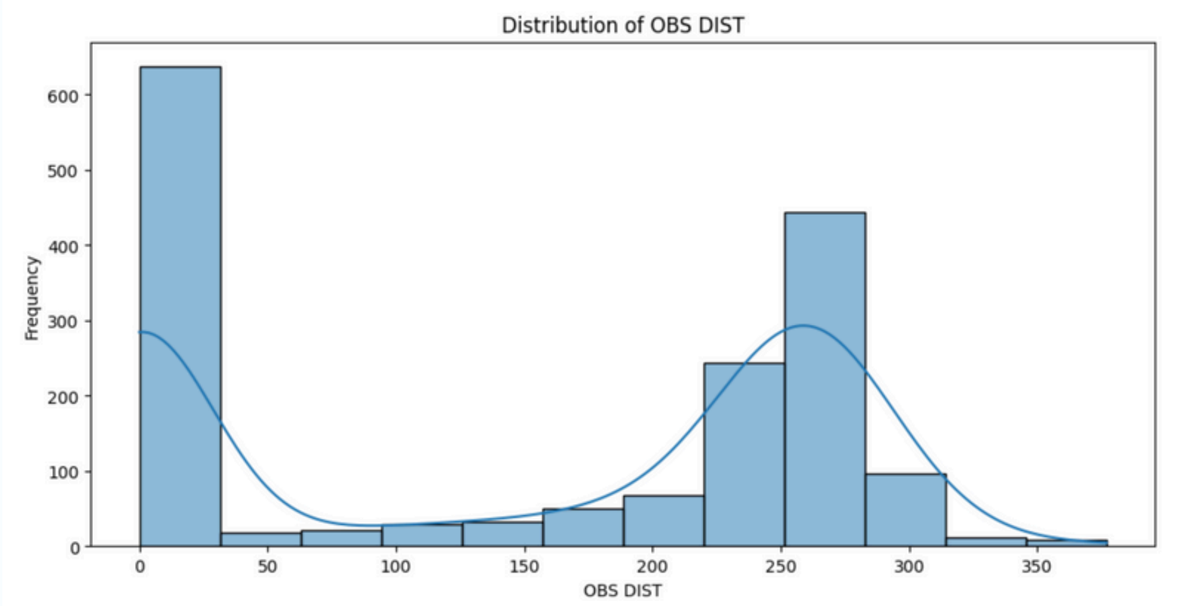
To show the distribution of CUR SPD

Using the histplot graph

To show the distribution of DRAFT FOR

EDA AND VISUALISATION





Using the histplot graph

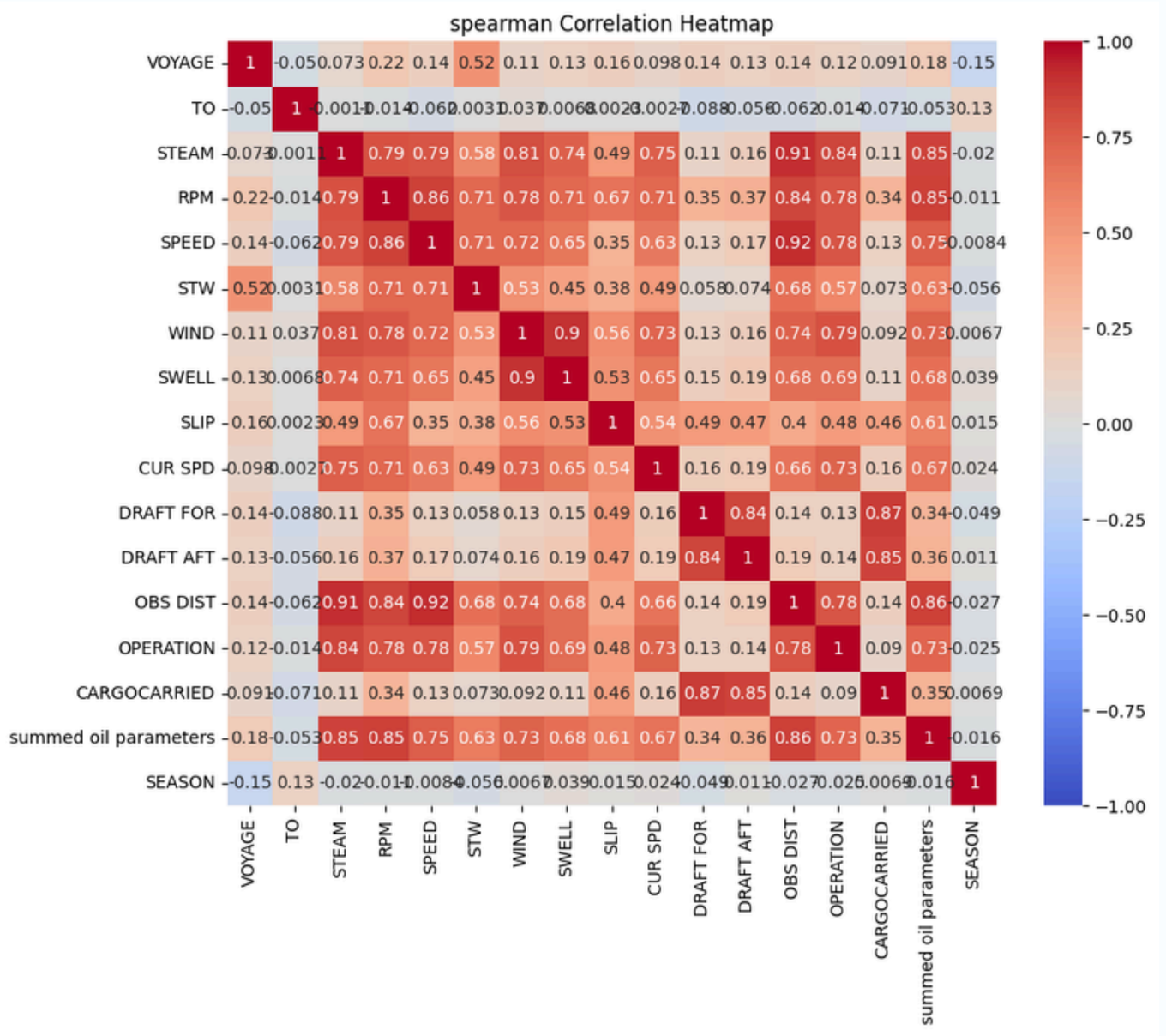
To show the distribution of DRAFT AFT

Using the histplot graph

To show the distribution of OBS DIST

CORRELATION

- In order to ease the processing power we used a heathmap graph to find a correlation between variables



FEATURES ENGINEERING

- First, we will sum all the oil consumption fields into summed oil parameters which is our target value
- Second , we convert the ‘date’ field into categorial ‘season’ field so we can use it for the prediction
- Observing the last stage (EDA), we will decide
  if any variable can affect the prediction.

MACHINE LEARNING

WHY CLASSIFACTION MODEL?

- **Nature of the Target Variable**:
  Our target variable is continuous value as it represent the total oil consumption ,
  a linear regression model is more appropriate
  since it directly predicts continuous value rather than a class membership.
- **Complex Relations between Variables:**
  In cases where the relationships between predictor variables and
  the target variable are complex and nonlinear,
  a linear regression model may have limitations.
  Linear regression is effective for capturing straightforward,
  linear relationships, but it may not fully capture complex patterns
  in the data without additional transformations
  or feature engineering.

MACHINE LEARNING

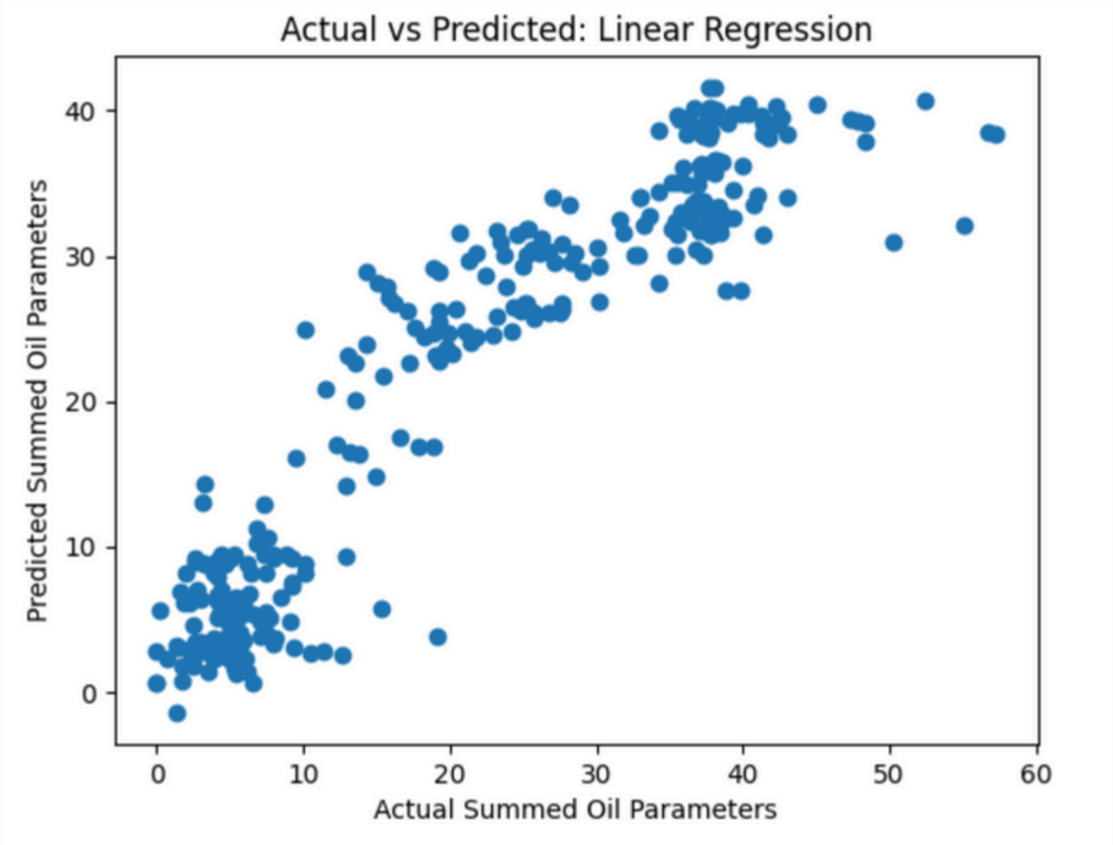
DATA PREPERATION

- Each categorical variable is converted into multiple binary variables, where each binary variable represents one category of the original variable. This process creates a new column for each category, with a value of 1 indicating the presence of that category and 0 otherwise. Dummy variables enable algorithms to interpret categorical data effectively and are essential for models that require numerical input.

MACHINE LEARNING
LINEAR REGRESSION

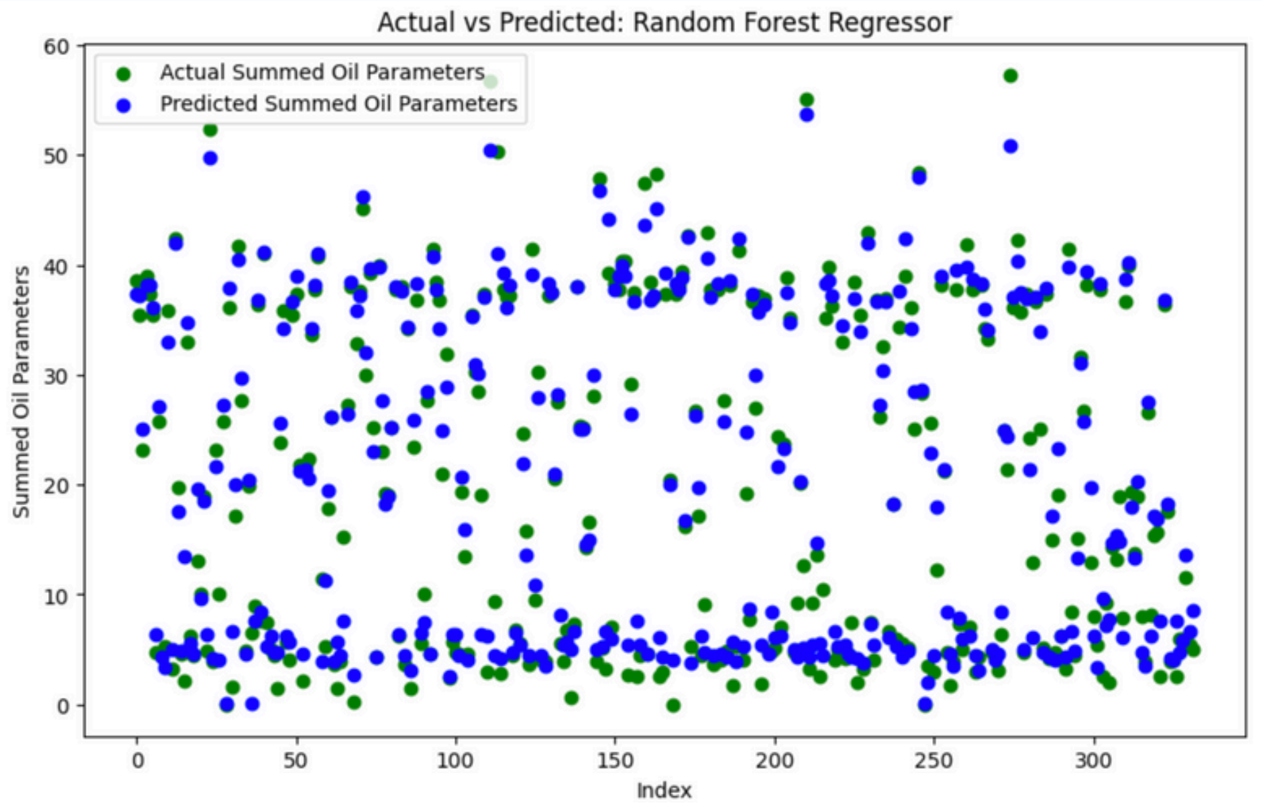
```
# Prepare the data for training
X = temp_encoded_df.drop(columns=['summed oil parameters'])
y = temp_encoded_df['summed oil parameters']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Standardize the features
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Train a Linear Regression model
model = LinearRegression()
model.fit(X_train_scaled, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test_scaled)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
print(f"R-squared: {r2}")
```

0.88



MACHINE LEARNING
RANDOM FOREST

```
# Prepare the data for training
X = my_scaled_data.drop(columns=['summed oil parameters'])
y = my_scaled_data['summed oil parameters']

# Select only columns that are of type float
x_encoded = pd.get_dummies(X)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(x_encoded, y, test_size=0.2,
random_state=42)

# Initialize and fit the RandomForestRegressor
rf_model = RandomForestRegressor()
rf_model.fit(X_train, y_train)

# Predict on the test set
y_pred = rf_model.predict(X_test)

# Evaluate the performance of the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
print(f"R^2 Score: {r2}")
```



0.971

SUMMARY AND CONCLUSION

The dataset contains many intricate relationships among features, with some features depending on or influencing others. We approached the problem as a regression task, aiming to model these relationships using a linear regression method. Despite the relatively small dataset, we achieved a strong fit, providing reliable predictions with a satisfying level of accuracy.

THANK YOU