Logistics Data Science Project Report

1. Introduction

This report details a data science project focused on analyzing and optimizing logistics operations using Python. The project covers data acquisition, exploratory data analysis (EDA), predictive modeling, and data visualization to derive actionable insights.

2. Data Acquisition and Setup

The dataset used for this project is the "Supply Chain Logistics Problem Dataset" obtained from Brunel University London via Figshare. The dataset contains various aspects of logistics operations, including order details, transportation information, and potential delays.

Project Structure

The project is organized into the following directories: - `data/`: Stores the raw and processed datasets. - `notebooks/`: Contains Python scripts for EDA, modeling, and visualization. - `reports/`: Stores generated reports, visualizations, and model evaluation results.

3. Exploratory Data Analysis (EDA)

EDA was performed to understand the dataset\'s structure, identify data types, check for missing values, and gain initial insights into the distribution of key variables. The dataset contains 9215 entries and 14 columns, including `Order ID`, `Order Date`, `Origin Port`, `Carrier`, `TPT` (Transportation Time), `Service Level`, `Ship ahead day count`, `Ship Late Day count`, `Customer`, `Product ID`, `Plant Code`, `Destination Port`, `Unit quantity`, and `Weight`.

Key Findings from EDA:

- The `Order Date` column was successfully parsed as datetime objects.
- Categorical features such as `Origin Port`, `Carrier`, `Service Level`, `Customer`, `Plant Code`, and `Destination Port` were identified as important for further analysis.
- Numerical features like `Unit quantity` and `Weight` show a wide range, indicating potential outliers or varied order sizes.

- The `Ship Late Day count` is a critical target variable, with most values being 0, indicating on-time deliveries, but some instances of delays.

4. Predictive Modeling and Optimization

To predict potential shipping delays, a RandomForestRegressor model was built. The `Ship Late Day count` was chosen as the target variable. Categorical features were one-hot encoded, and irrelevant columns were dropped.

Model Evaluation Results:

```
 Mean Absolute Error (MAE): 0.04
Mean Squared Error (MSE): 0.07
Root Mean Squared Error (RMSE): 0.27
R-squared (R2): 0.49
```

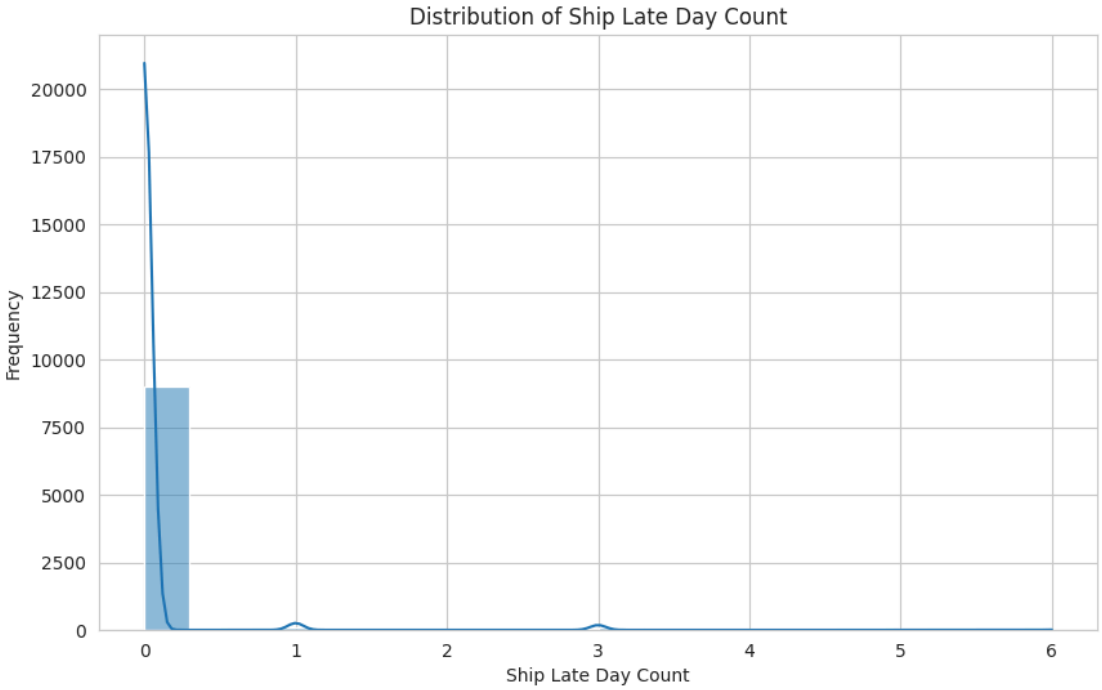
The model achieved a relatively low Mean Absolute Error (MAE) of 0.04, indicating that on average, the predictions for `Ship Late Day count` are very close to the actual values. The R-squared (R2) value of 0.49 suggests that approximately 49% of the variance in `Ship Late Day count` can be explained by the model. While this is a reasonable starting point, further feature engineering and model tuning could improve performance.

5. Comprehensive Visualizations

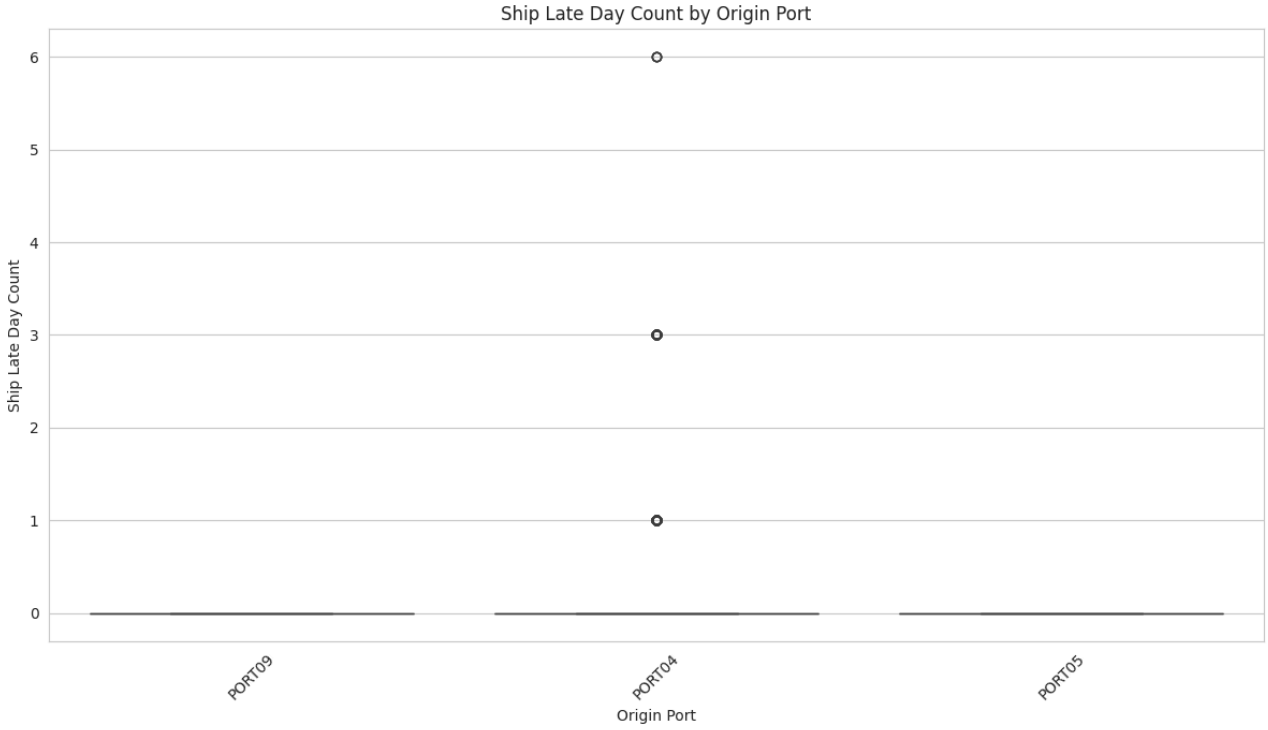
Several visualizations were generated to provide a deeper understanding of the data and the factors influencing shipping delays. These visualizations are saved as PNG files in the `reports/` directory.

Visualizations include:

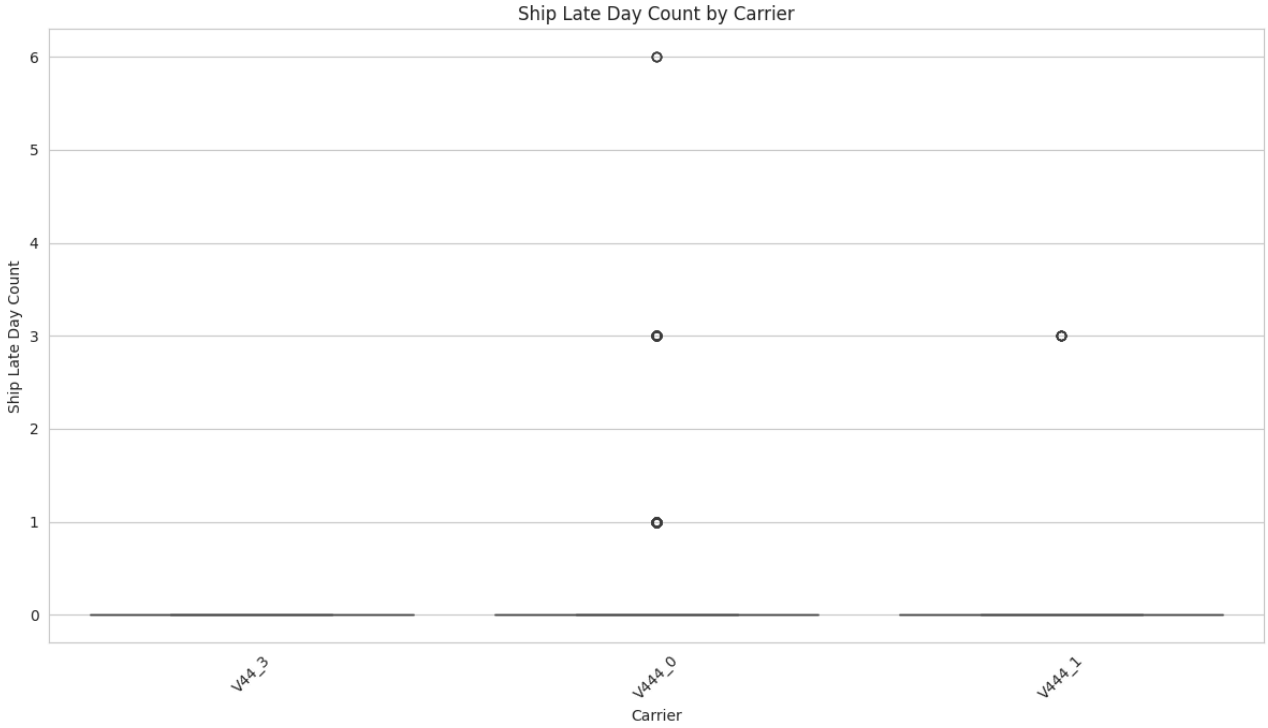
Distribution of Ship Late Day Count



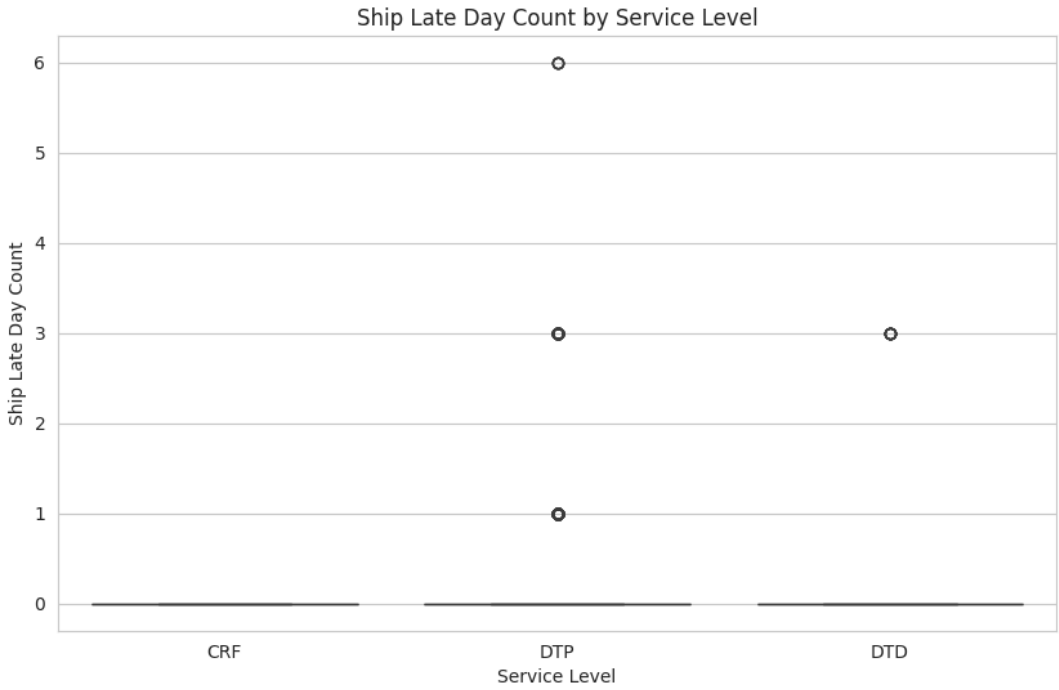
Ship Late Day Count by Origin Port



Ship Late Day Count by Carrier



Ship Late Day Count by Service Level



These visualizations help identify patterns and potential areas for improvement in logistics operations. For example, certain origin ports or carriers might consistently experience higher delays, indicating a need for targeted interventions.

6. Conclusion and Recommendations

This project demonstrates the application of data science techniques to analyze and optimize logistics operations. By leveraging historical data, we can identify factors contributing to shipping delays and build predictive models to forecast them.

Recommendations:

- **Focus on problematic areas:** Investigate origin ports and carriers identified in the visualizations as having higher delays to understand underlying causes and implement corrective actions.
- **Refine predictive model:** Explore additional features, such as external factors (weather, traffic) or more granular time-based features, to improve the accuracy of the predictive model. Consider advanced modeling techniques like time series analysis if daily or hourly data becomes available.
- **Implement real-time monitoring:** Integrate the predictive model into a real-time monitoring system to provide early warnings for potential delays, allowing for proactive intervention.
- **Optimize routing:** Utilize the insights from delay predictions to optimize routing and scheduling, potentially reducing transportation costs and improving delivery times.

References

[1] Supply Chain Logistics Problem Dataset. Figshare. Available at: https://brunel.figshare.com/articles/dataset/Supply_Chain_Logistics_Problem_Dataset/7558679