# K-Means Clustering for Supply Chain Delay Segmentation

## 1. Introduction

This project presents an unsupervised learning approach to identify hidden patterns behind logistics delays in a smart supply chain environment. By applying K-Means Clustering, we uncover operational groupings that support delay diagnostics and process optimization in real-world logistics systems.

While using the k-means clustering algorithm, the first step is to indicate the number of clusters (k) that we wish to produce in the final output. The algorithm begins by randomly selecting k objects from the dataset to serve as the initial centers for our clusters. These selected objects are the cluster means, also known as centroids. Then, the remaining objects have an assignment of the closest centroid. This centroid is defined by the Euclidean Distance present between the object and the cluster mean. We refer to this step as "cluster assignment". When the assignment is complete, the algorithm proceeds to calculate the new mean value of each cluster present in the data. After recalculating the centers, the observations are checked to see if they are closer to a different cluster. Using the updated cluster means, the objects undergo reassignment. This goes on repeatedly through several iterations until the cluster assignments stop altering. The clusters that are present in the current iteration are the same as the ones obtained in the previous iteration.

## 2. Objective

This project applies K-Means Clustering on a smart logistics dataset to segment supply chain operations based on shipment and delay characteristics. The primary goal is to uncover hidden patterns and groupings that explain logistical delays, enabling actionable decisions for supply chain optimization.

## 3. Dataset Description

The Smart Logistics Supply Chain dataset was chosen due to its real-world applicability, richness in operational variables, and relevance to modern, IoT-integrated logistics systems. It includes key performance indicators such as waiting time, inventory levels, traffic status, environmental conditions, and delivery delays — all essential for understanding the root causes of inefficiencies. The combination of sensor-driven and categorical data makes it ideal for applying clustering to discover meaningful operational segments that could inform decision-making in supply chain management.

- Source: https://www.kaggle.com/datasets/ziya07/smart-logistics-supply-chain-dataset/data

## 4. Methodology and Analysis:

### 4.1. Data Cleaning

- Removed rows with missing values in key features.
- Encoded categorical features (Traffic_Status, Logistics_Delay) using label encoding.

### 4.2. Feature Selection & Scaling

Selected numeric and encoded categorical features:

- Waiting_Time, Inventory_Level, Traffic_Status, Temperature, Humidity, Logistics_Delay

Standardized all selected features using StandardScaler.

### 4.3. Clustering with K-Means

- Elbow Method used to determine optimal clusters (k=3)

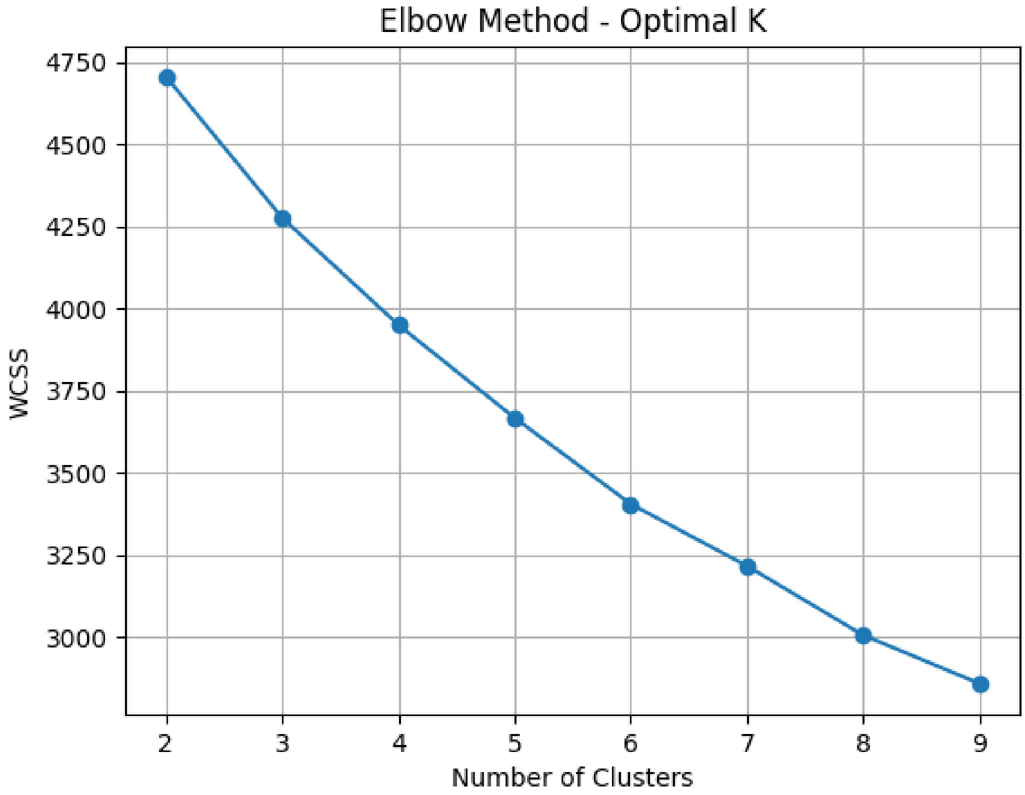
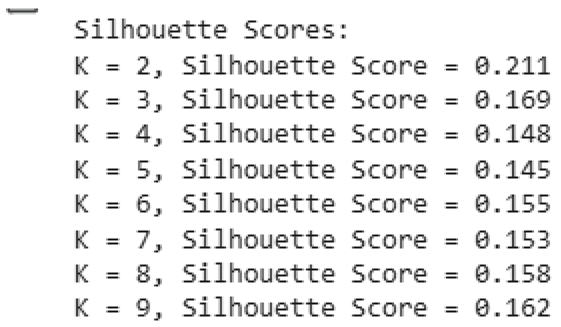


*Figure 1 Elbow Method showing the optimal number of clusters based on within-cluster sum of squares (WCSS).*

- Silhouette Score verified compactness and separability



```
Silhouette Scores:
K = 2, Silhouette Score = 0.211
K = 3, Silhouette Score = 0.169
K = 4, Silhouette Score = 0.148
K = 5, Silhouette Score = 0.145
K = 6, Silhouette Score = 0.155
K = 7, Silhouette Score = 0.153
K = 8, Silhouette Score = 0.158
K = 9, Silhouette Score = 0.162
```

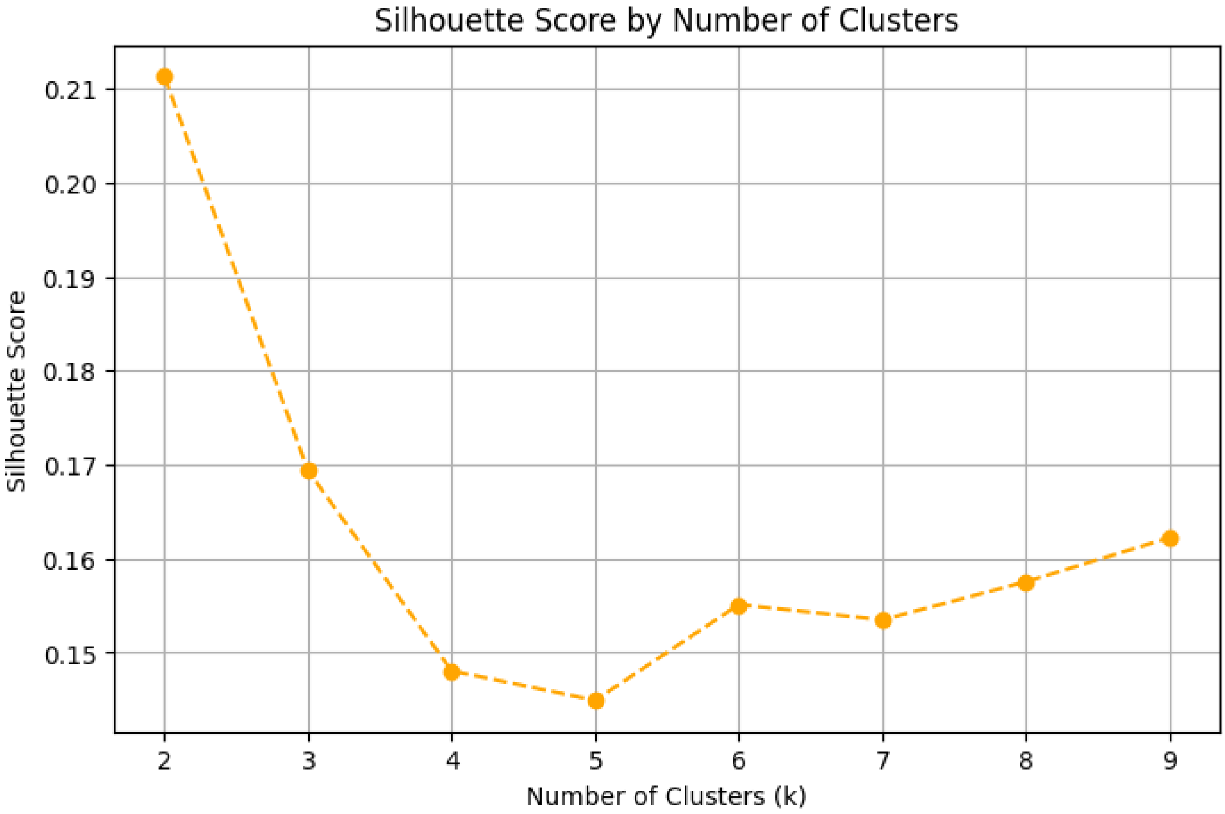


*Figure 2 Silhouette Score plot to evaluate clustering compactness across different values of k.*

### 4.4. Dimensionality Reduction

- Used PCA to project high-dimensional clusters into 2D for visualization

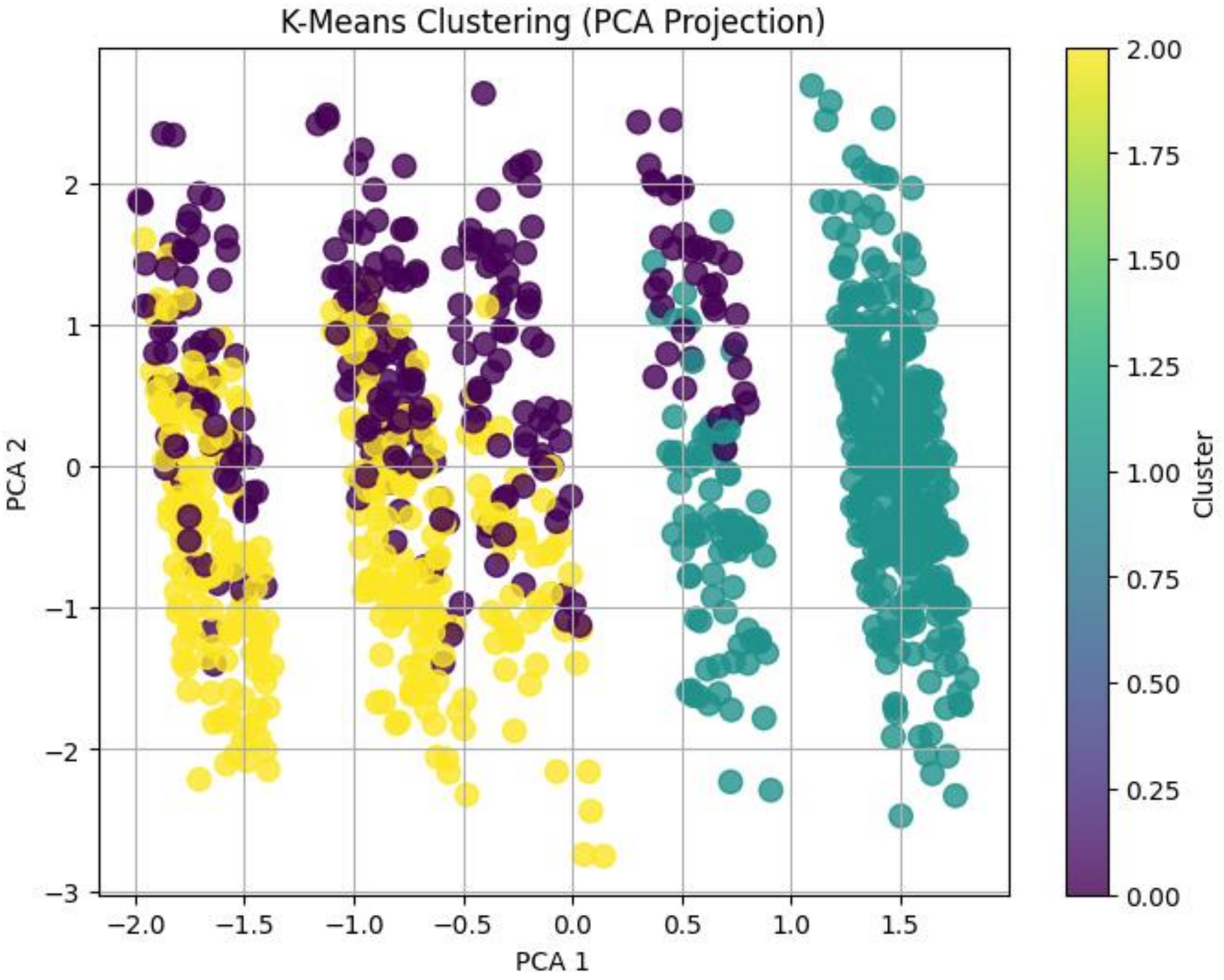
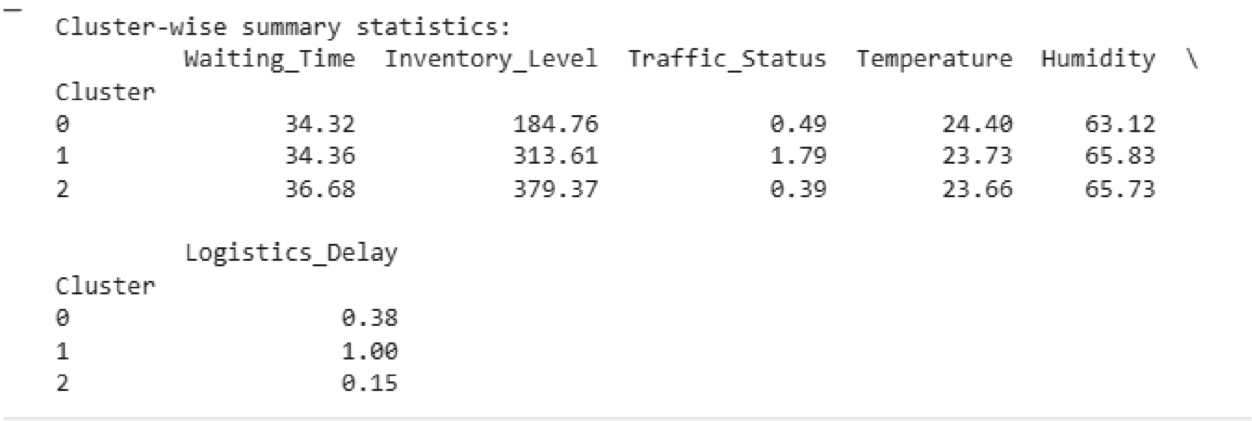


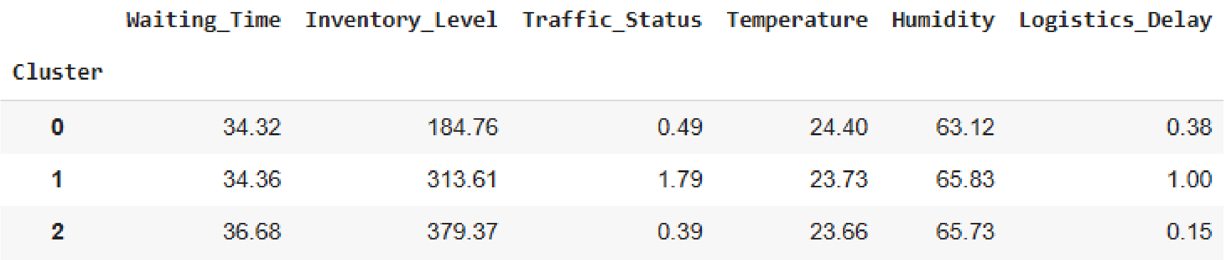
*Figure 3 PCA 2D projection of K-Means clusters for visualizing separability in reduced dimensions.*

## Cluster Summary



```
Cluster-wise summary statistics:
         Waiting_Time  Inventory_Level  Traffic_Status  Temperature  Humidity  \
Cluster
0               34.32           184.76            0.49        24.40     63.12
1               34.36           313.61            1.79        23.73     65.83
2               36.68           379.37            0.39        23.66     65.73

         Logistics_Delay
Cluster
0                   0.38
1                   1.00
2                   0.15
```



| Cluster | Waiting_Time | Inventory_Level | Traffic_Status | Temperature | Humidity | Logistics_Delay |
|---|---|---|---|---|---|---|
| **0** | 34.32 | 184.76 | 0.49 | 24.40 | 63.12 | 0.38 |
| **1** | 34.36 | 313.61 | 1.79 | 23.73 | 65.83 | 1.00 |
| **2** | 36.68 | 379.37 | 0.39 | 23.66 | 65.73 | 0.15 |

## 5. Visualizations

### 5.1. Radar Chart

Compares average feature values across clusters, revealing:

- Cluster 1: High traffic, high delays
- Cluster 2: High inventory, low delays
- Cluster 0: Low inventory, moderate delays

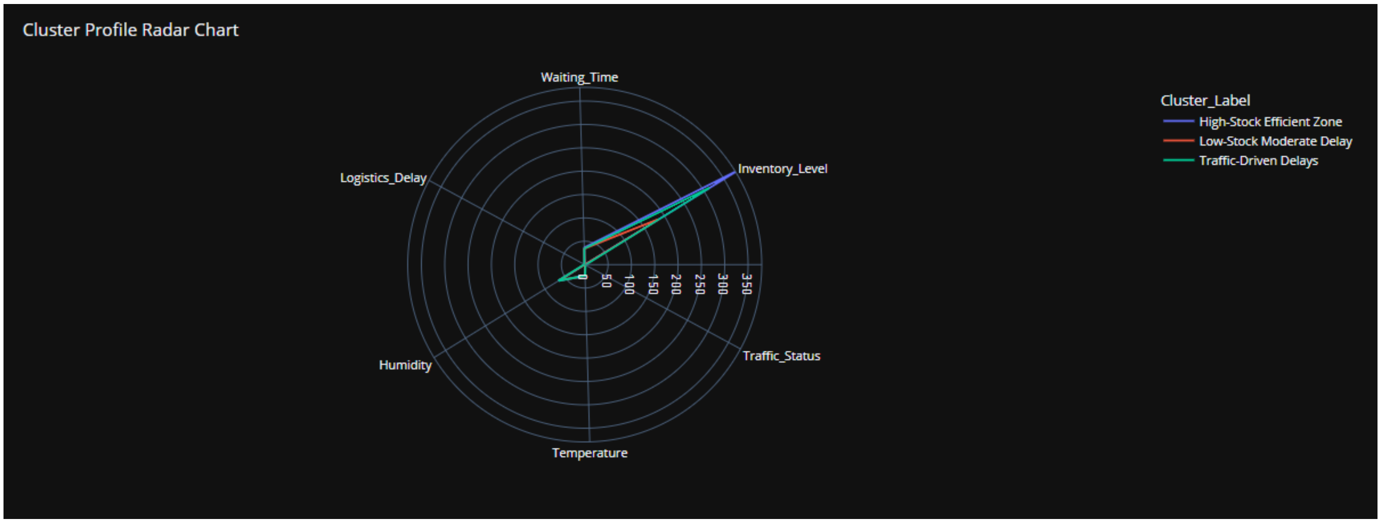


*Figure 4 Radar Chart comparing average feature values across clusters, showing distinct operational profiles.*

## 5.2. Boxplots by Cluster

Shows variation and outliers across clusters for all six features.

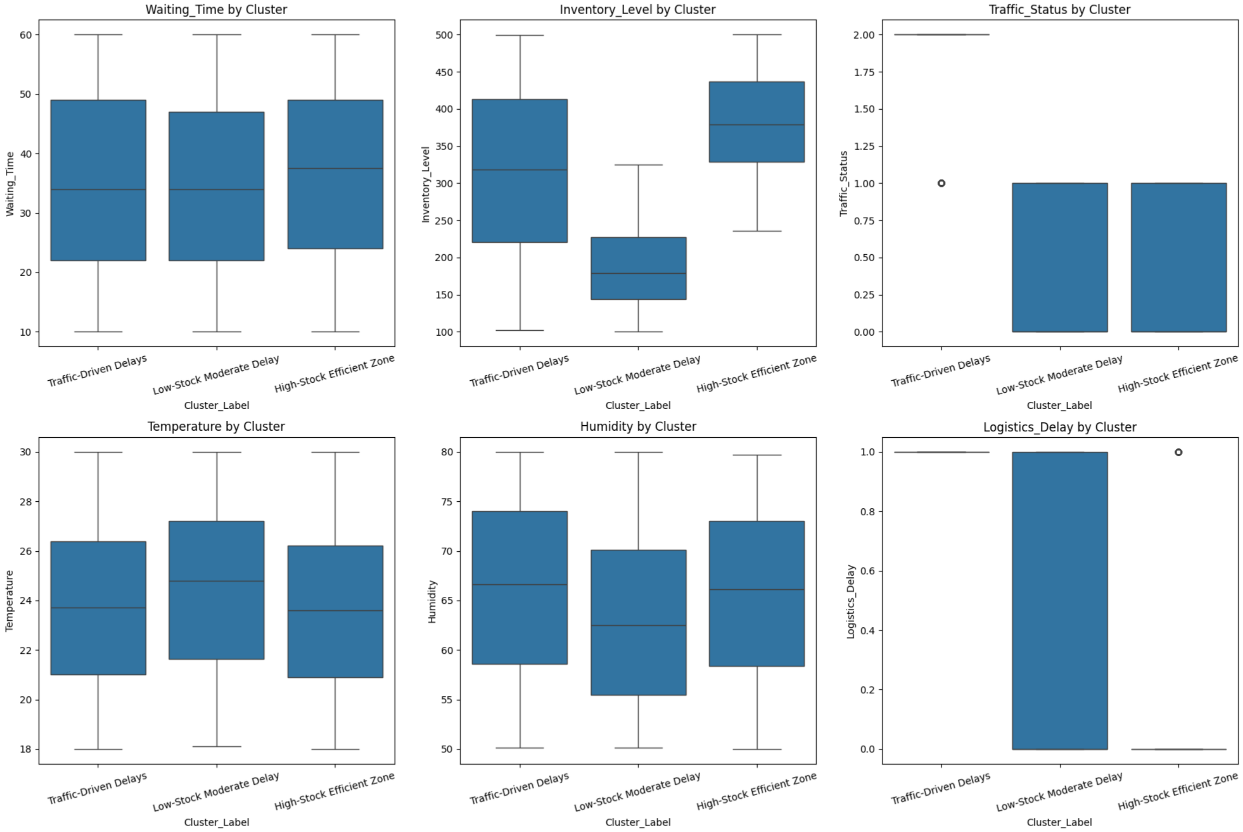
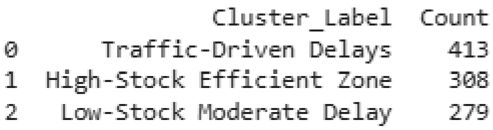


*Figure 5 Boxplots*



```
               Cluster_Label  Count
0      Traffic-Driven Delays    413
1  High-Stock Efficient Zone    308
2   Low-Stock Moderate Delay    279
```

## 5.3. PairPlot

Explored inter-feature relationships, showed natural group separations in clusters.

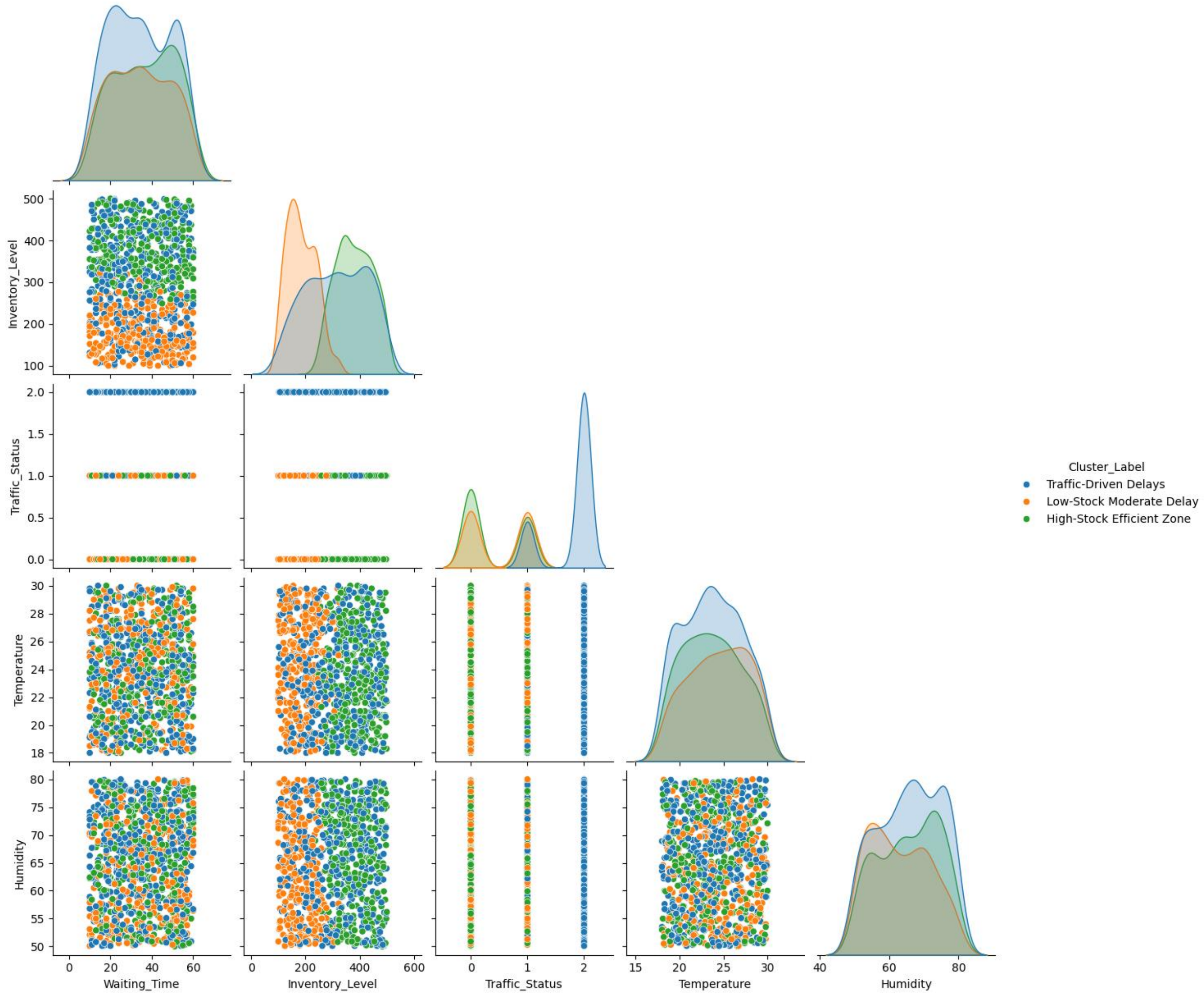


*Figure 6 PairPlot visualizing feature relationships and natural cluster separability in high-dimensional space.*

## 6. Final Cluster Summary

| Cluster Label | Key Traits | Suggested Action |
|---|---|---|
| Low-Stock Moderate Delay | Low inventory, average waiting, low traffic, mid delay | Increase buffer stock |
| Traffic-Driven Delays | High traffic, 100% delay, mid inventory, medium humidity | Optimize routing and delivery timing |
| High-Stock Efficient Zone | High inventory, low traffic, low delay, longer waiting time | Maintain SOP and use as a benchmark |

## 7. Business Insights

- Cluster 0 likely represents rural or less-optimized routes with moderate delays due to limited stock.
- Cluster 1 signals urban bottlenecks where traffic is the key root cause.
- Cluster 2 is an operationally efficient segment with sufficient inventory and minimal delays.

## 8. Conclusion

This project successfully demonstrates how K-Means clustering can be applied to real-world logistics data to uncover meaningful operational patterns. The identified segments can now be used to drive targeted decisions, optimize delivery operations, and reduce supply chain delays.

*Created by Afrasyab Khan* | *https://github.com/iAfrasyab* |
*https://www.linkedin.com/in/afrasyabkhan/*