Project by :
Elin Nurulita

Shoe Warehouse Operational Efficiency Analysis

Data cleaning with Python, Exploratory with SQL and Visualization with Looker Studio

Oktober 2025

Elin Nurulita

PROJECT OVERVIEW

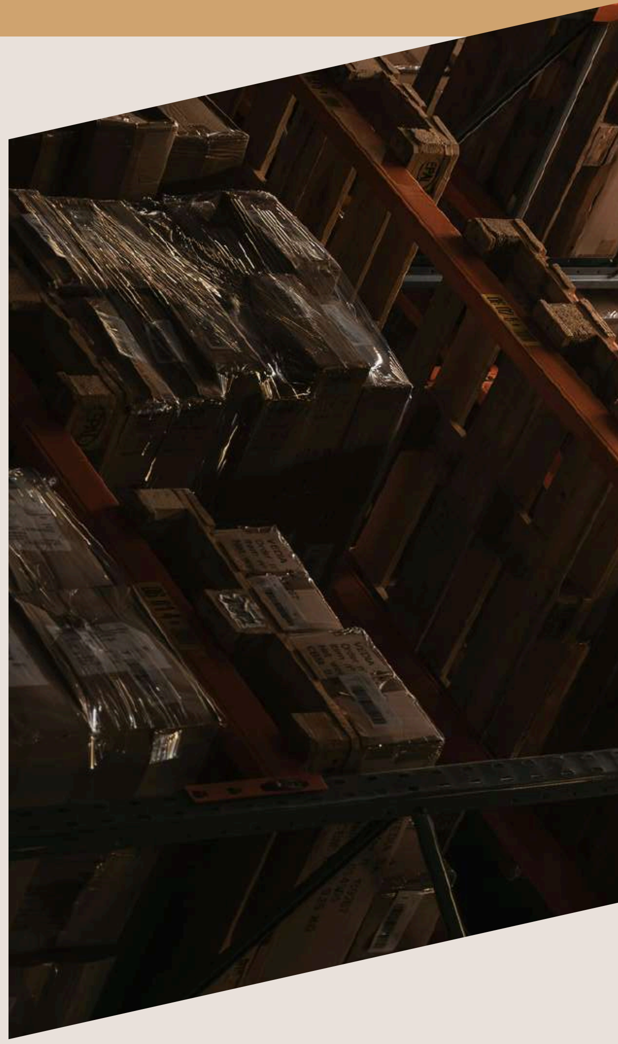
OBJECTIVE:

- Clean the warehouse operation dataset to ensure data accuracy, consistency and realibility for further analysis
- Perform EDA to identify insights and opportunies for optimizing the performance and efficiency of the operation
- Develop an interactive dashboard in Looker Studio

TOOLS:

- Python (Google Colab)
- BigQuery
- Looker Studio

DATASET DESCRIPTION



DATASET	: picking data sepatu
SIZE	: 50.000 rows, 15 columns
COLUMNS	: warehouse_zone, date, shift, picker_id, error_type, shipping_method, quality_check_passed, destination_country, etc.

DATA CLEANING

1. LOAD AND INSPECT

- The dataset is loaded using the pandas library with the read_csv() function
- Inspect the dataset to understand its data types, sample rows, and identify any potential issus such as missing values or data inconsistencies.

a. Load the dataset and inspect the dataset info (data types and sample rows)

```python
import pandas as pd

# Load dataset
df = pd.read_csv('picking_data_sepatu.csv')

# Column types and non null counts
df.info()
df.head()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50250 entries, 0 to 50249
Data columns (total 15 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   date                50250 non-null  object
 1   shift               50250 non-null  int64
```

b. Check missing values of the dataset

```python
# Check number of missing values per column
print("\n=== Missing Values per Column ===")
print(df.isna().sum())
```

```
=== Missing Values per Column ===
date                    0
shift                   0
order_id                0
picker_id               0
warehouse_zone          0
brand                   0
product_category        0
item_count            402
picking_time_min      606
destination_country     0
status                  0
error_type          48735
shipping_method         0
supervisor_id           0
quality_check_passed    0
dtype: int64
```

1. LOAD AND INSPECT

c. Finding inconsistency in dataset

```python
print("\n=== Unique warehouse zones ===")
print(df['warehouse_zone'].unique())
```

```
=== Numeric Columns ===
Index(['shift', 'item_count', 'picking_time_min'], dtype='object')

=== Object Columns ===
Index(['date', 'order_id', 'picker_id', 'warehouse_zone', 'brand',
       'product_category', 'destination_country', 'status', 'error_type',
       'shipping_method', 'supervisor_id'],
      dtype='object')

=== Sample order_id inconsistencies (floats as IDs) ===
   date  shift  order_id  picker_id  warehouse_zone  brand  product_categor

=== Unique brand values (potential inconsistencies) ===
['Puma' 'Adidas' 'LocalBrand' 'Nike' 'Reebok' 'Converse' 'Adidas '
 'Converse ' 'adidas' 'converse' 'reebok' 'nike' 'puma' 'Nike '
 'localbrand' 'Reebok ' 'Puma ']

=== Unique warehouse zones ===
['D' 'C' 'A' 'B' 'E' 'b' 'e' 'c' 'a' 'd']
```

2. HANDLE MISSING VALUES

- Missing values were found in the **error_type**, **item_count**, and **picking_time_min** columns.
- To handle them, missing **error_type** values were replaced with "No Error", **item_count** was filled with the median, and **picking_time_min** with the mean. This ensures data consistency and reduces bias in the analysis.

```
# Handle missing values
df['error_type'] = df['error_type'].fillna('No Error'
df['item_count'] = df['item_count'].fillna(df['item_c
df['picking_time_min'] = df['picking_time_min'].filln

print("\n=== Missing Values per Column ===")
print(df.isna().sum())
```

```
=== Missing Values per Column ===
date                  0
shift                 0
order_id              0
picker_id             0
warehouse_zone        0
brand                 0
product_category      0
item_count            0
picking_time_min      0
destination_country   0
status                0
error_type            0
shipping_method       0
supervisor_id         0
quality_check_passed  0
dtype: int64
```

3. FIX DATA TYPES, STANDARDIZE AND NORMALIZE

- Converted the ID columns (**order_id, picker_id, and supervisor_id**) to string type
- Normalized by removing extra spaces and standardizing letter cases for text columns (**brand, warehouse_zone, status, product_category, etc**) for example, ‘ adidas’ -> ‘Adidas and ‘b’ → ‘B’

```
=== Data Types ===
date                    datetime64[ns]
shift                            int64
order_id                        object
picker_id                       object
warehouse_zone                  object
brand                           object
product_category                object
item_count                     float64
picking_time_min               float64
destination_country             object
status                          object
error_type                      object
shipping_method                 object
supervisor_id                   object
quality_check_passed              bool
dtype: object

=== Sample Cleaned Data ===
```

	date	shift	order_id	picker_id	warehouse_zone
0	2025-04-18	1	ORD-001641	P02	D
1	2025-	2	ORD-	P07	C

4. REMOVE DUPLICATES

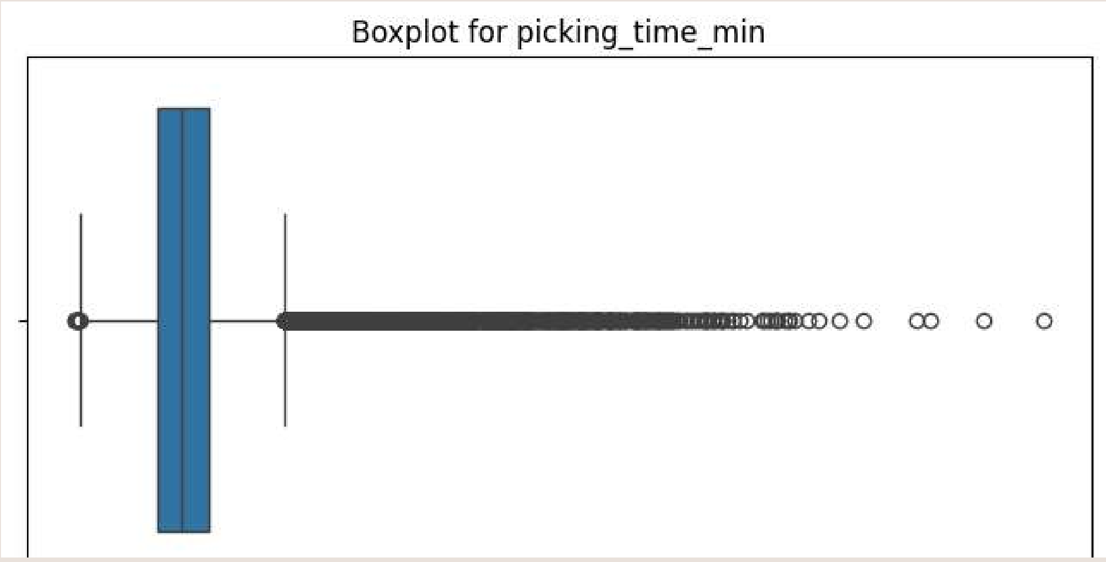
- Removed the duplicates using pandas : drop_duplicates()

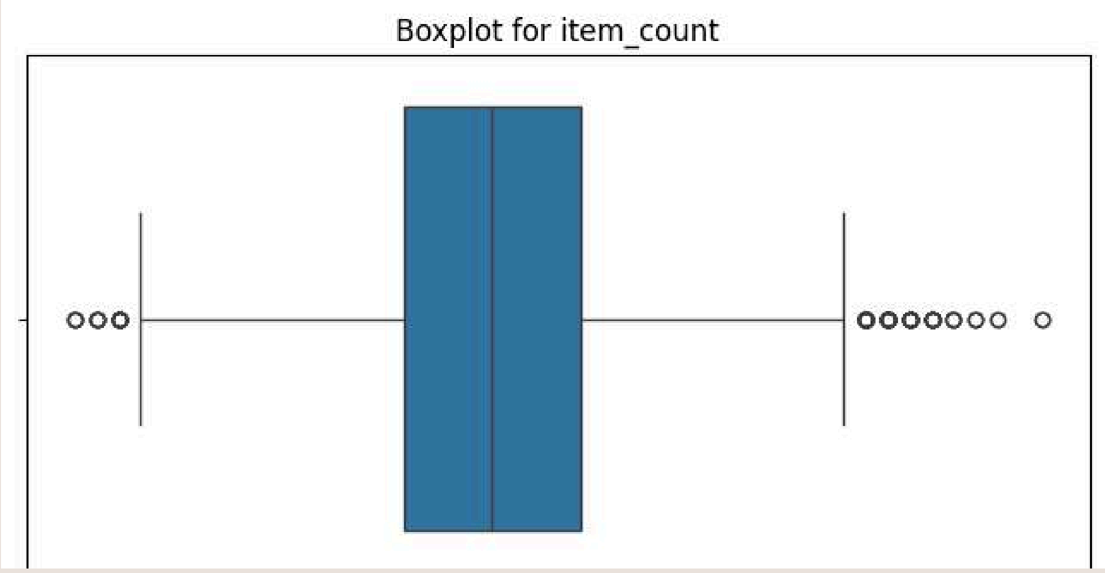
```
# Remove duplicate
df.drop_duplicates(inplace=True)
```

Result :

```
Initial rows: 50250
Duplicate rows found: 250
Final rows after removing duplicates: 50000
```

5. HANDLE OUTLIERS





Handling Outliers (Winsorization using IQR method)

- Outliers in the **item_count** and **picking_time_min** columns were treated using the **Winsorization** technique based on the **Interquartile Range (IQR)**.
- Values below the lower limit or above the upper limit were capped (replaced) with the respective boundary values using the clip() function in pandas.

6. EXPORT DATASET

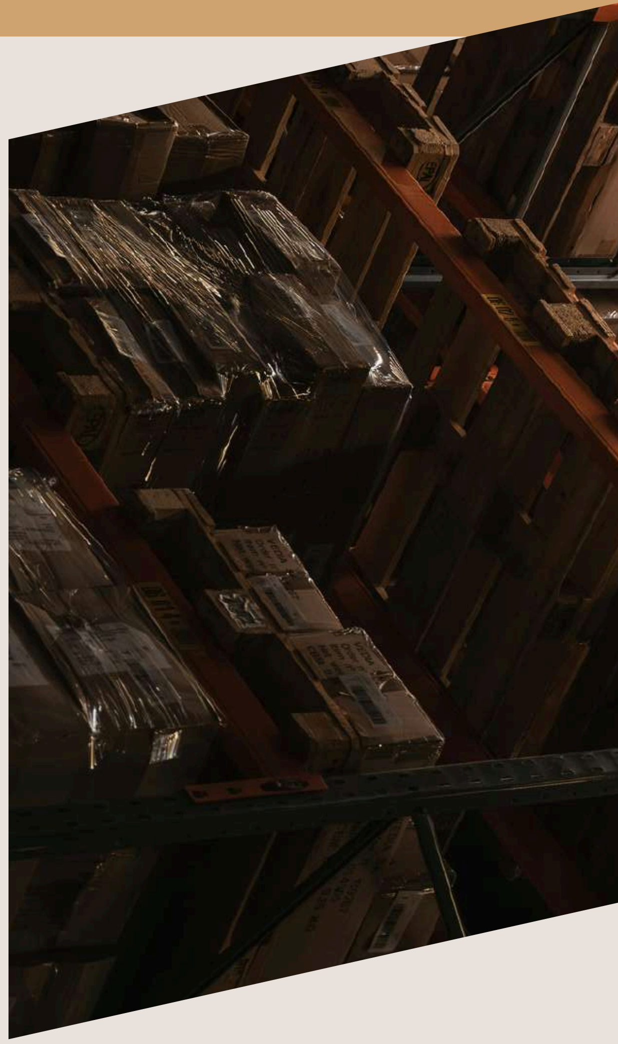
- Exported dataset into CSV & Excel
- Dataset ready for exploration and visualisation

```
# Save as CSV
df.to_csv('/content/picking_clean.csv', index=False)

print("✅ File saved as picking_clean.csv")
```

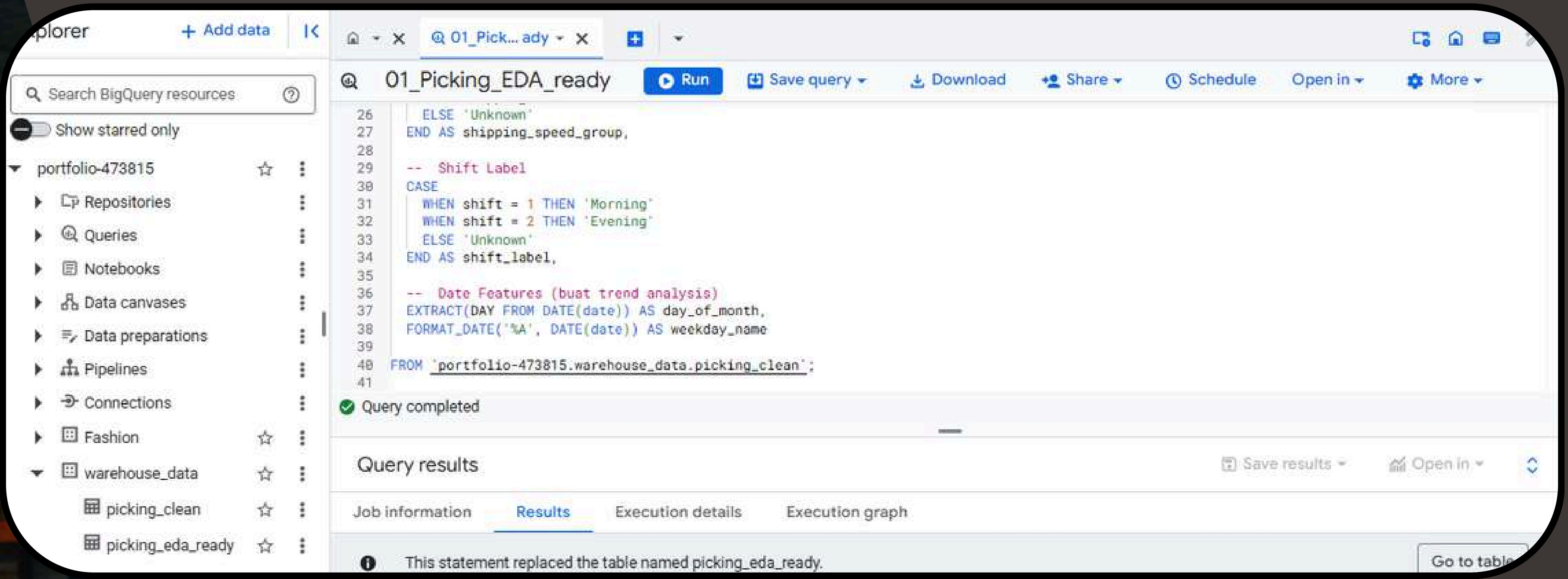
```
✅ File saved as picking_clean.csv
```

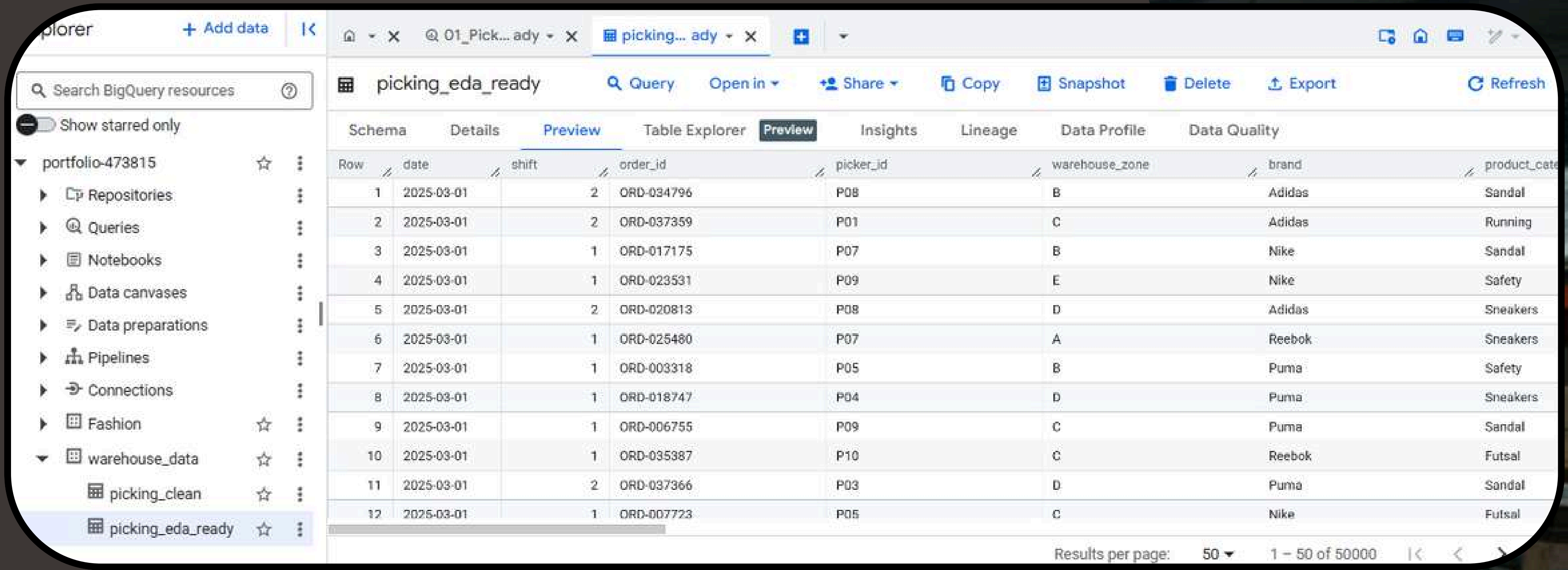
```
# Save as Excel
df.to_excel('/content/picking_clean.xlsx', index=False)

print("✅ File saved as picking_clean.xlsx")
```

```
✅ File saved as picking_clean.xlsx
```



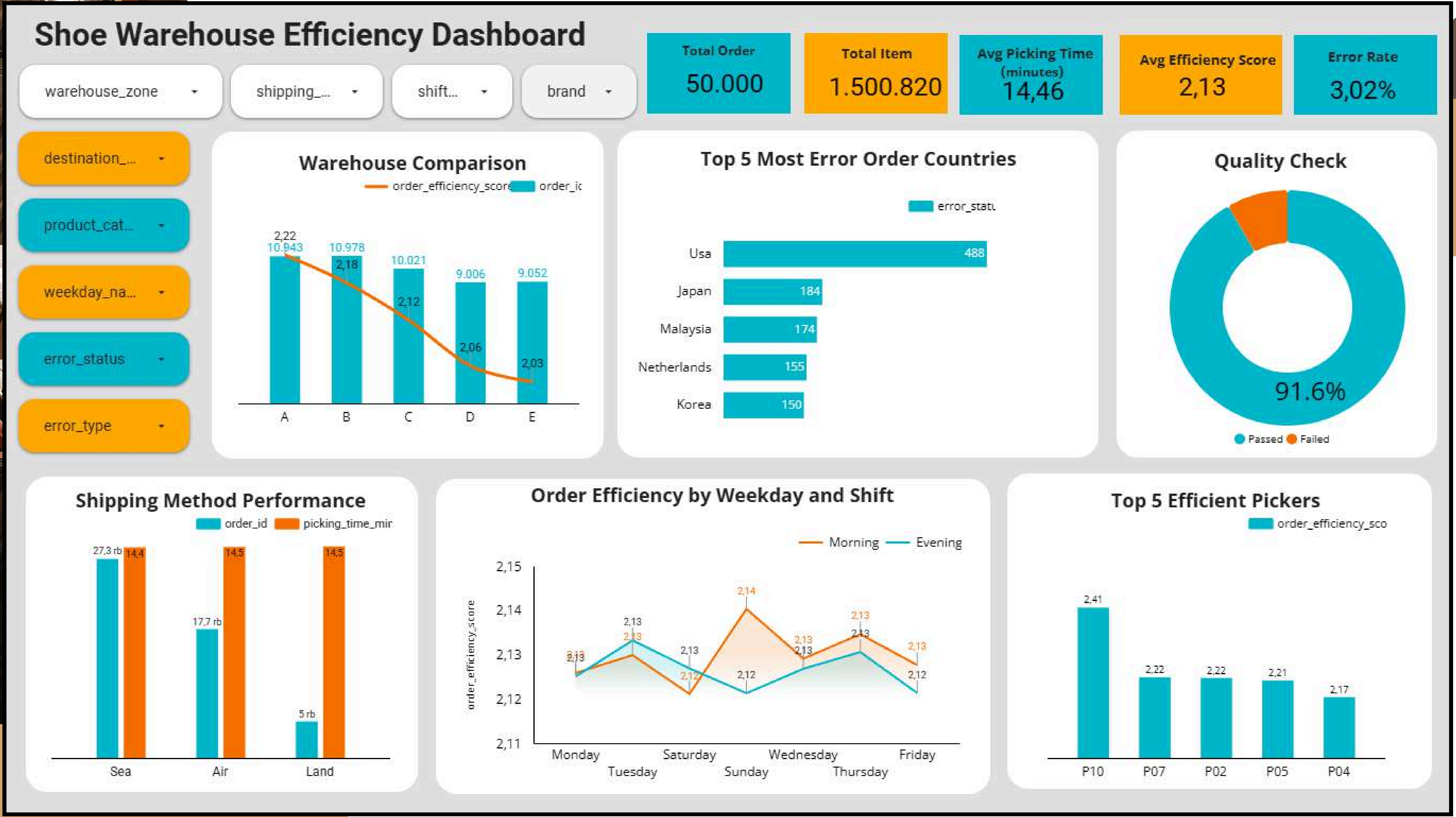
DATA EXPLORATION





- After the data was cleaned using Python, it was uploaded to Google BigQuery for further data exploration and basic querying. BigQuery was used to efficiently process and analyze large datasets using SQL.
- The cleaned and processed data was then connected to Looker Studio to create interactive dashboards and visualizations.

OVERALL DASHBOARD



PERFORMANCE METRICS

Total Order	Total Item	Avg Picking Time (minutes)	Avg Efficiency Score	Error Rate
50.000	1.500.820	14,46	2,13	3,02%

- A total of 50,000 orders were recorded, containing 1,500,820 items.
- The average picking time per order is 14.46 minutes, showing a moderate operational pace.
- The average efficiency score is 2.13 items per minute, indicating consistent performance across warehouse zones.
- The overall error rate is 3.02%, meaning that around 3 out of every 100 orders experience an error.

Total item and Picking time Correlation

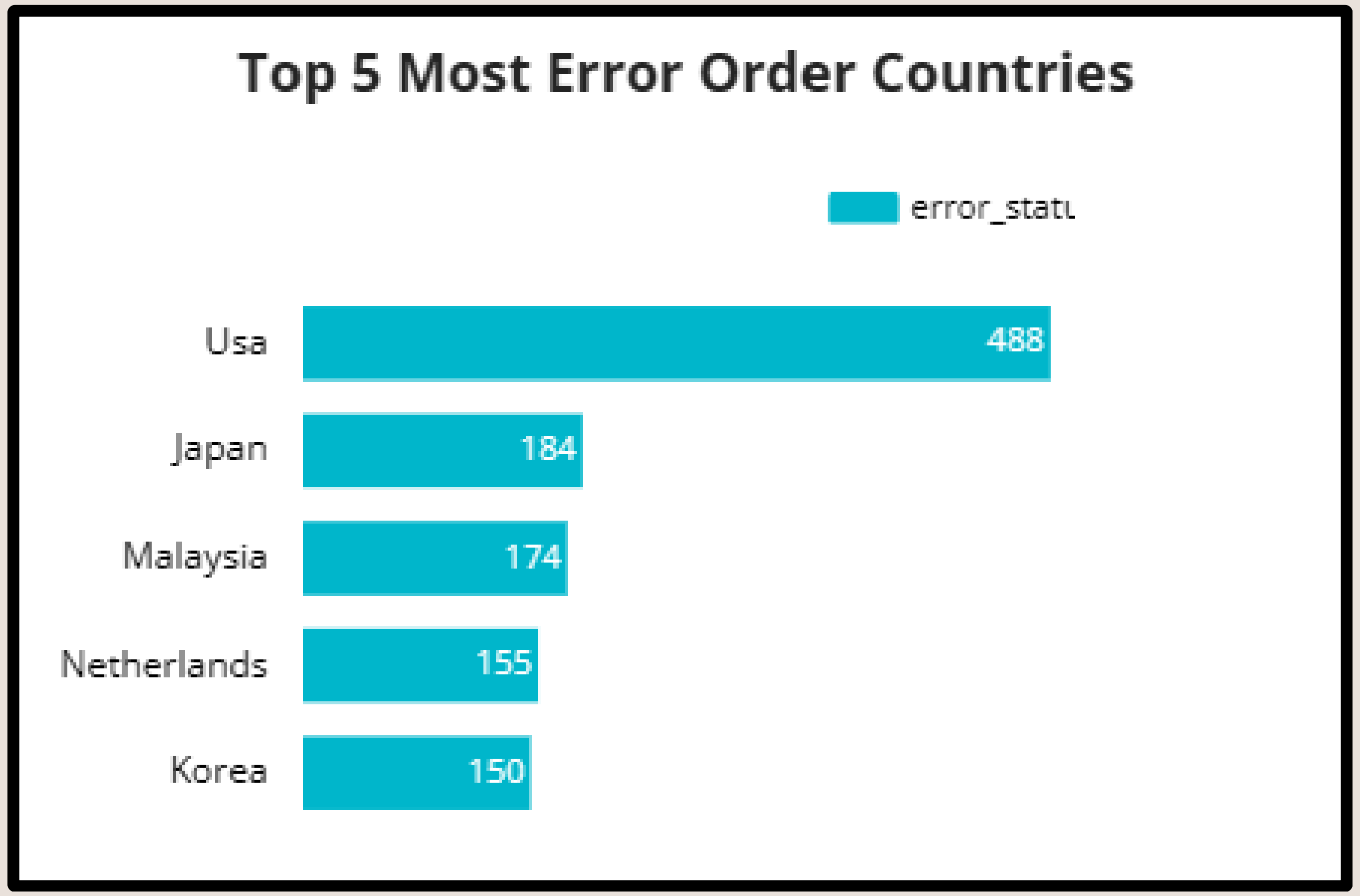
Row	warehouse_zone	corr_item_picking	avg_picking_time	avg_item_count
1	D	0.729	14.92	29.98
2	B	0.722	14.15	30.07
3	A	0.692	13.84	29.96
4	C	0.729	14.48	30.03
5	E	0.734	15.11	30.06

Correlation (0.69–0.73) → **strong positive relationship**
More items = longer picking time, consistently across all zones.
This shows order complexity drives picking duration

Recommendations :

- Apply batch/zone picking for high-item orders.
- Improve item mapping to reduce walking time.

MOST ERROR ORDER COUNTRY



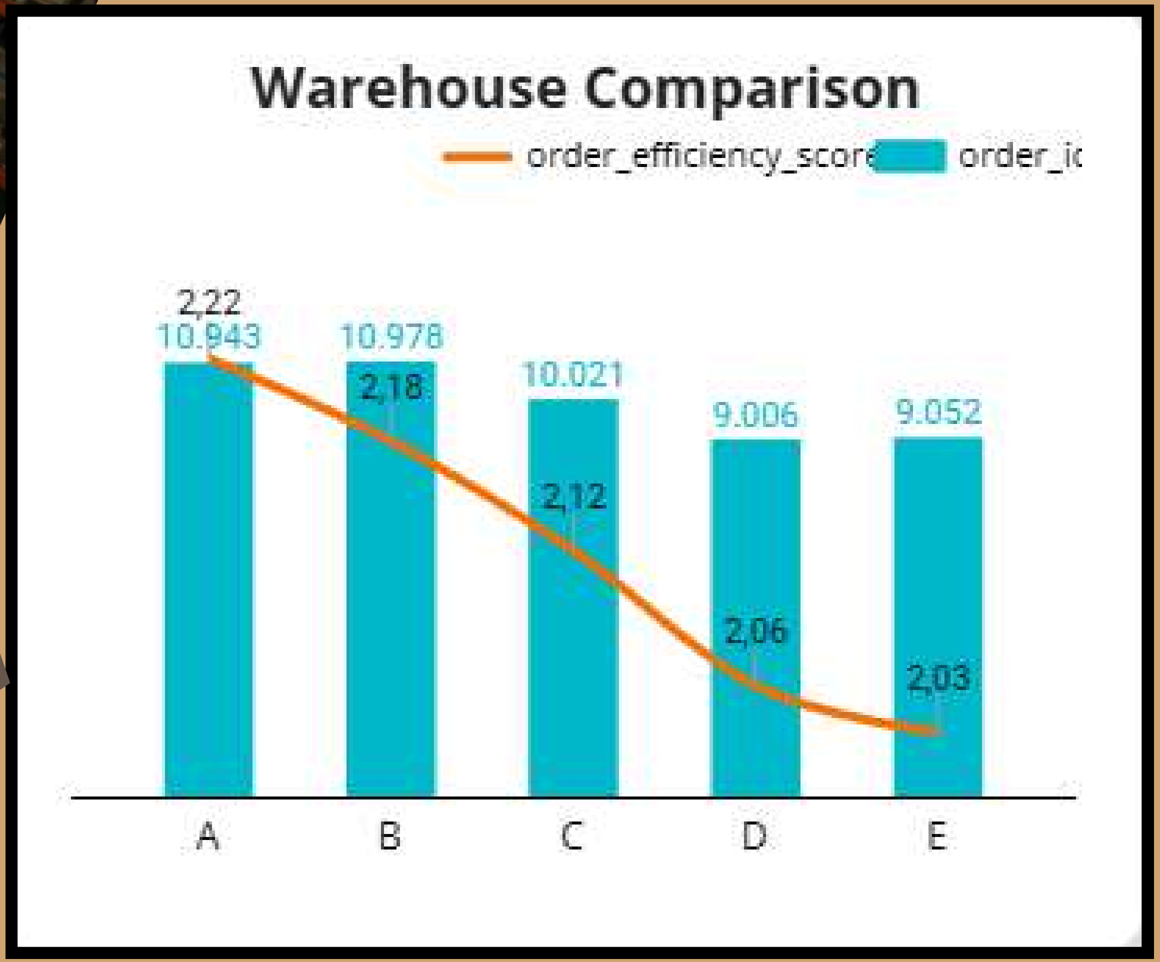
error_type	total_error	error_percentage...
Misspick	175	35.86
Wrong Label	165	33.81
Delay	148	30.33

- **USA has the highest error rate — 32% (488 orders)** of total errors. Indicates frequent operational inefficiencies in orders shipped to the USA.
- Top error types:
  - **Misspick** — 35%
  - **Wrong Label** — 33%
  - **Delay** — 30%
- Similar proportions across error types → suggests **systemic consistency issues** (picking, labeling, and shipping).
- Next step: Identify **which warehouse** is most involved in **US-bound orders** to locate the **root cause**.

Next chart: Warehouse performance comparison (efficiency vs total order).

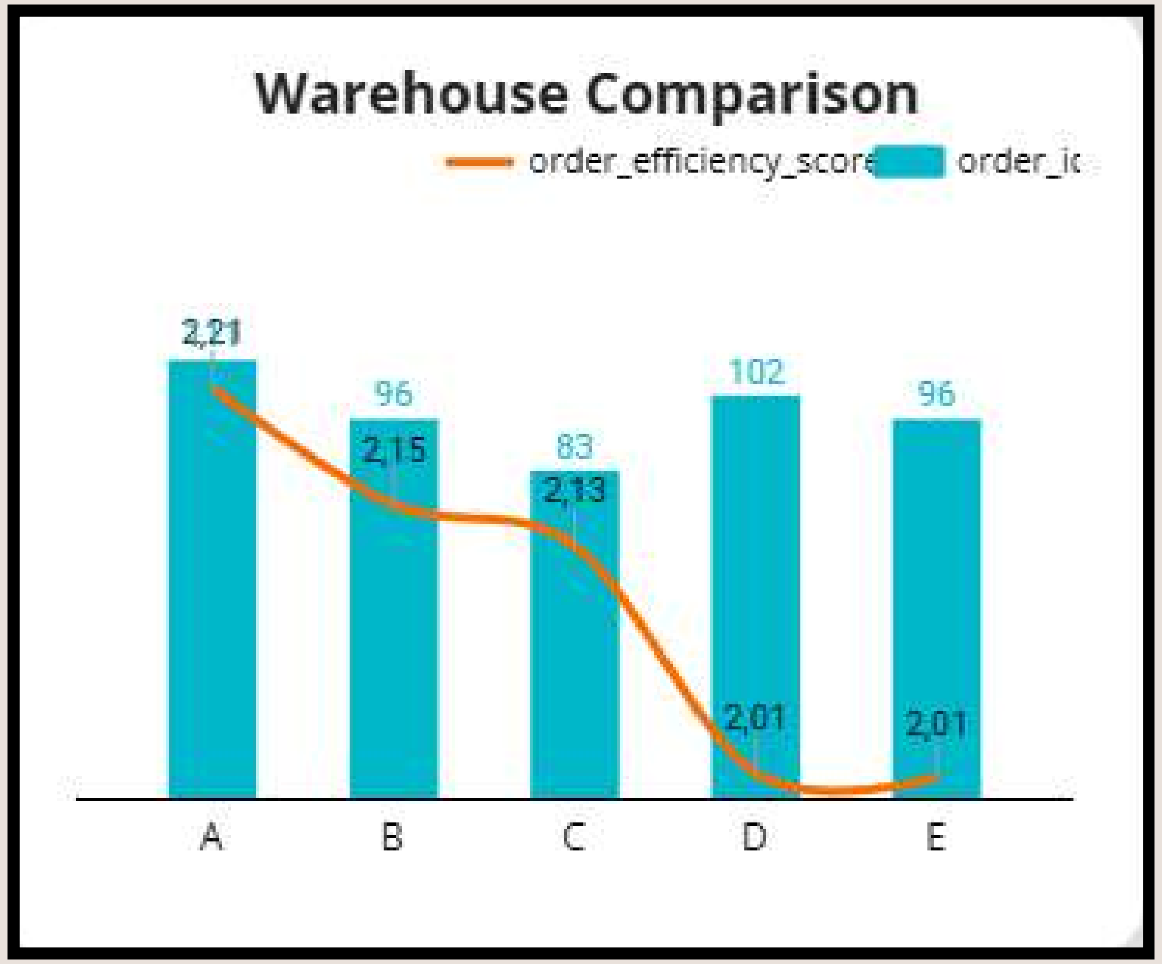
WAREHOUSE PERFORMANCE COMPARISON

Overall Performance



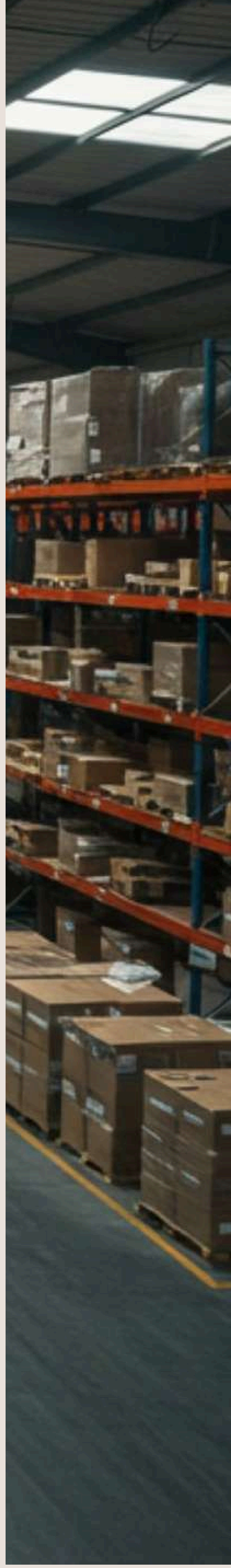
- Warehouse A & B → highest order volume and top efficiency.
- Warehouse D & E → lowest volume and lowest efficiency.

Filter : USA, Error Order



For **USA Error Orders**:

- Warehouse A → handles most US error orders, but still efficient (2.21) → sign of overload risk.
- Warehouse D & E → also handle many US error orders, low efficiency → need process improvement.

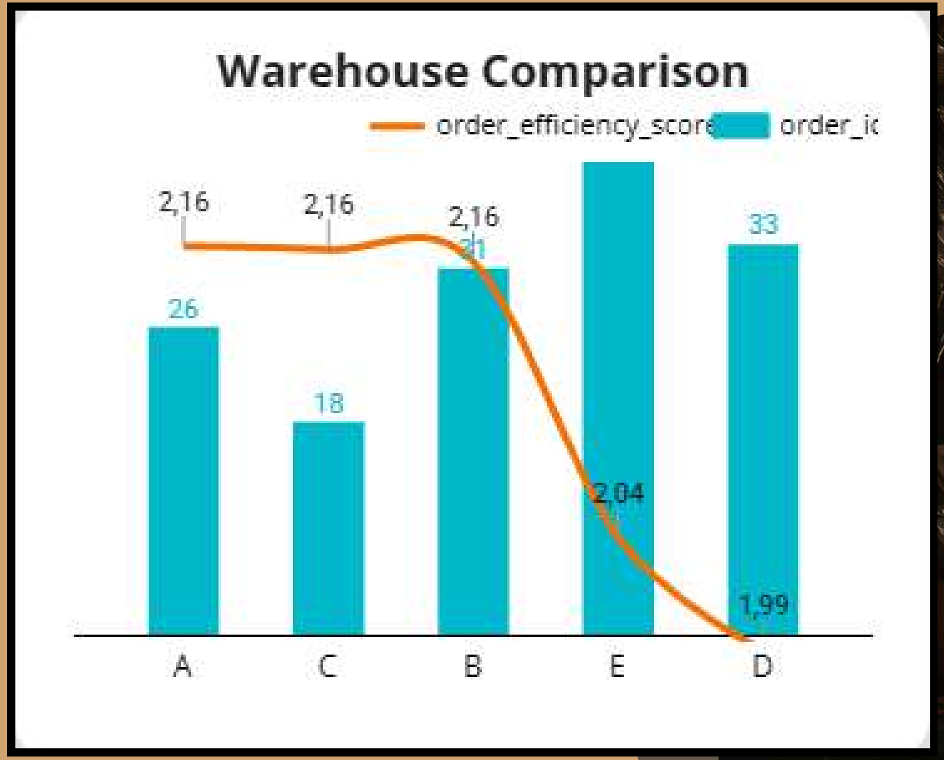


WAREHOUSE PERFORMANCE COMPARISON (ERROR BREAKDOWN)

Filter : USA, Error Order, Misspick

Filter : USA, Error Order, Wrong Label

Filter : USA, Error Order, Delay



Error Breakdown:

- **Misspick** : A, D, E
- **Wrong Label** : A, D, B, C
- **Delay** : E, D

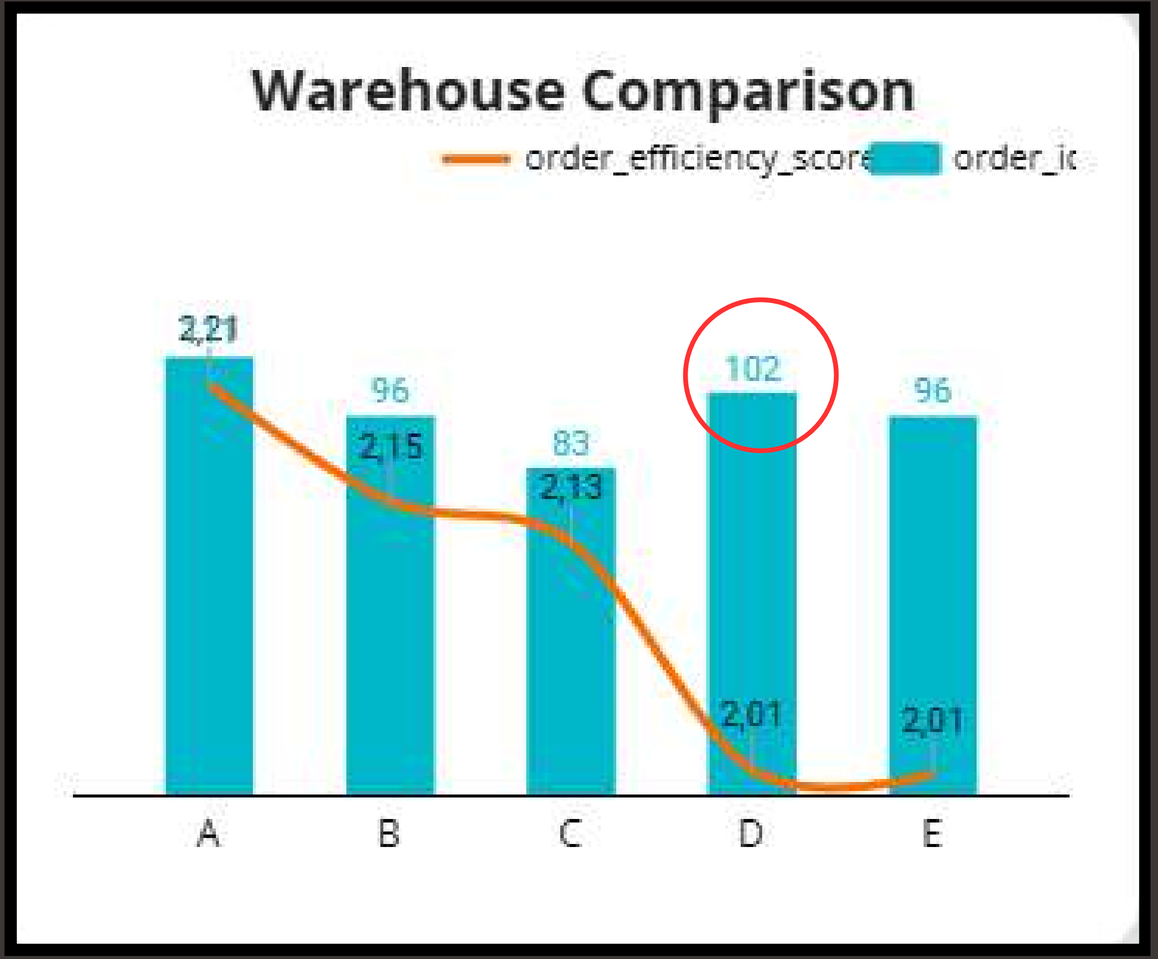
Insights & Recommendations :

- Warehouse A shows strong efficiency under heavy load → monitor for potential capacity strain.
- Redistribute order volume (especially US shipments) from A to D/E to balance workloads.
- Use A's workflow as a benchmark to improve D & E processes.
- Monitor efficiency trends — if A's rate drops, scale capacity or automate.

While Warehouse A performs well despite heavy workload, Warehouse D stands out for the opposite reason — fewer orders but significantly lower efficiency. This makes Warehouse D a key focus point for deeper investigation.

DEEP DIVE : WAREHOUSE D (USA ERROR ORDER)

Warehouse Comparison (USA + Error order)

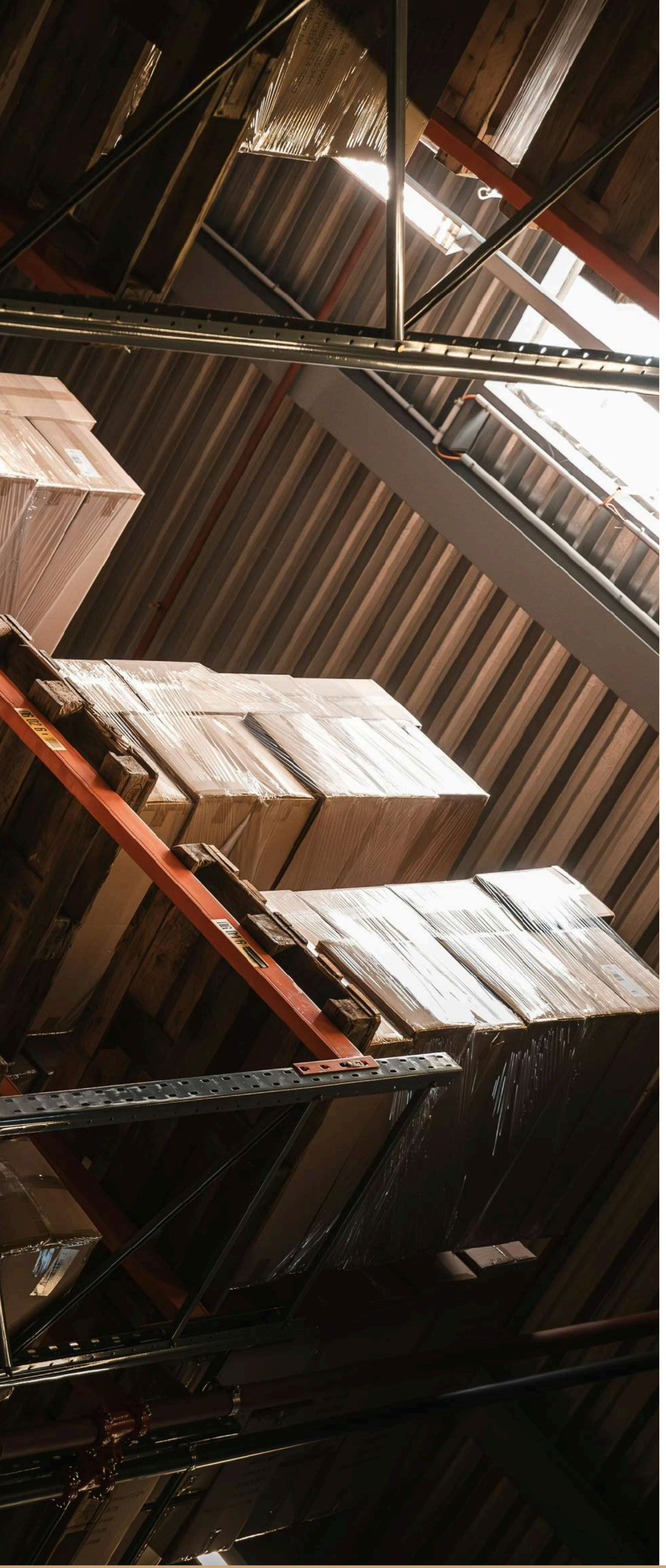


Error Distribution Accross Warehouse D (USA + Error Order)

error_type	total_error	error_percentage...
Wrong Label	35	34.31
Misspick	34	33.33
Delay	33	32.35

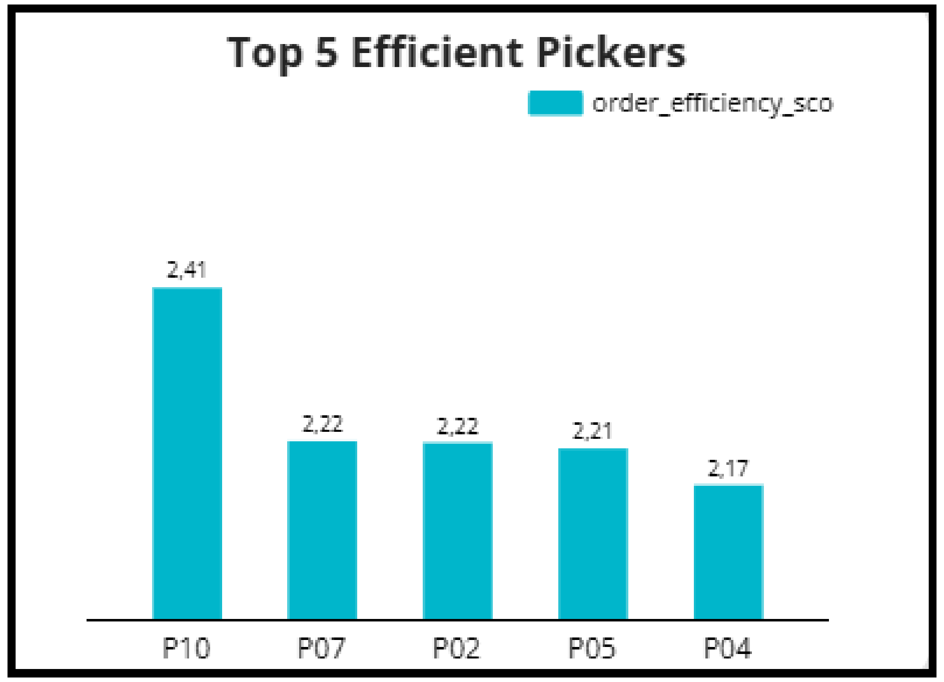
- Warehouse D handle many USA error orders while the efficiency is low
- Error distribution : Wrong Label 34% | Misspick 33% | Delay 32%
- Meaning : Error occur evenly (systemic inefficiency, not single process issue
- Efficiency : low (possible workflow and QC issues

Since Warehouse D shows consistently low efficiency across all error types, the next step is to analyze picker-level performance to determine whether these issues are linked to specific individuals or general workflow problems.



PICKER PERFORMANCE OVERVIEW

Avg Picking Time (minutes)	Avg Efficiency Score	Error Rate
14,46	2,13	3,02%



Error Percentage of Pickers

picker_id	efficiency_score	percentage_error
P10	2.41	2.82
P07	2.22	3.32
P02	2.22	3.09
P05	2.21	3.38
P04	2.17	2.58
P01	2.14	3.04
P09	2.07	2.76
P03	2.03	2.82
P08	1.92	3.19
P06	1.89	3.2

← lowest efficiency picker

- Top 5 most efficient pickers: P10, P07, P02, P05, P04 → avg. efficiency 2.13 items/min, all above warehouse average.
- Error rate among top pickers remains moderate (2.6–3.4%), showing strong balance between speed and accuracy.
- The least efficient picker (P06) has 1.89 items/min with slightly higher error rate (3.2%) → potential for training improvement.

Higher efficiency generally aligns with lower error rates, but there's still a small gap between top and low performers that indicates room for process standardization.

DEEP DIVE : PICKERS PERFROMANCE IN WAREOUSE D (USA + ERROR ORDER)

Avg Picking Time (minutes)	Avg Efficiency Score
15,37	2,01

Avg Picking Time and Total Order (focus : Error order)

picker_id	avg_picking_time	total_error_order
P07	14.05	13
P08	16.65	13
P06	16.45	13
P03	17.78	11
P01	14.93	10
P02	14.96	10
P05	14.06	9
P04	15.53	9
P09	15.14	7
P10	12.61	7

Efficiency and Error Percentage

picker_id	efficiency_score	percentage_error
P10	2.29	2.47
P07	2.17	4.66
P05	2.17	3.32
P02	2.15	3.57
P04	2.13	3.09
P01	2.08	3.57
P09	1.99	2.41
P03	1.96	3.73
P08	1.87	4.3
P06	1.83	4.23

- Avg. efficiency 2.01 items/min, avg. picking time 15.37 min → slower than overall average (14.46 min).
- Pickers handling most USA error orders: P07, P08, P06, P03, P01 with higher error rates (3.5–4.6%) and lower efficiency (≈1.8–2.1 items/min).
- Top efficient pickers (P10, P07, P02, P04, P05) handle fewer errors → shows uneven task distribution.

Error orders in Warehouse D are mainly handled by lower-efficiency pickers, suggesting both workload imbalance and quality control issues during error-prone tasks.

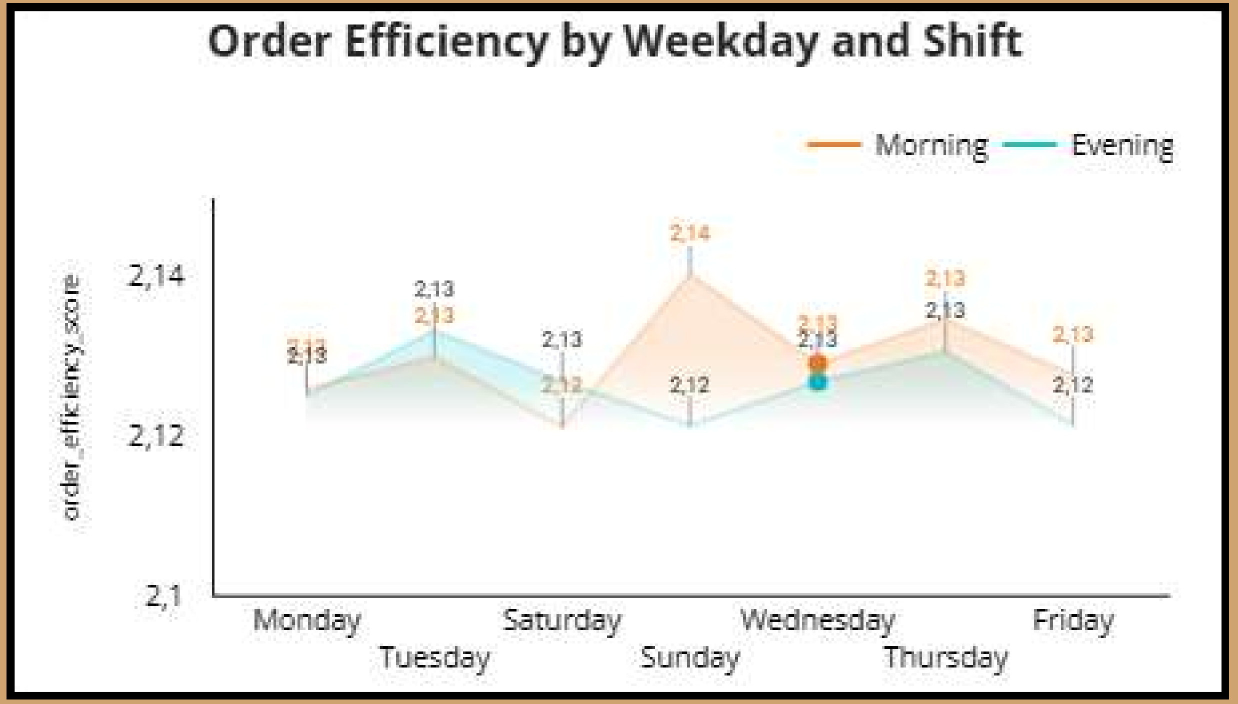
Recommendation:

- Redistribute complex/US-bound orders to balance workload.
- Provide targeted training for low-efficiency pickers (P06 & P08).
- Review picking workflow to reduce handling time spikes (>15 min)

Following picker-level analysis, the next step focuses on efficiency trends by weekday & shift

WEEKDAY AND SHIFT EFFICIENCY

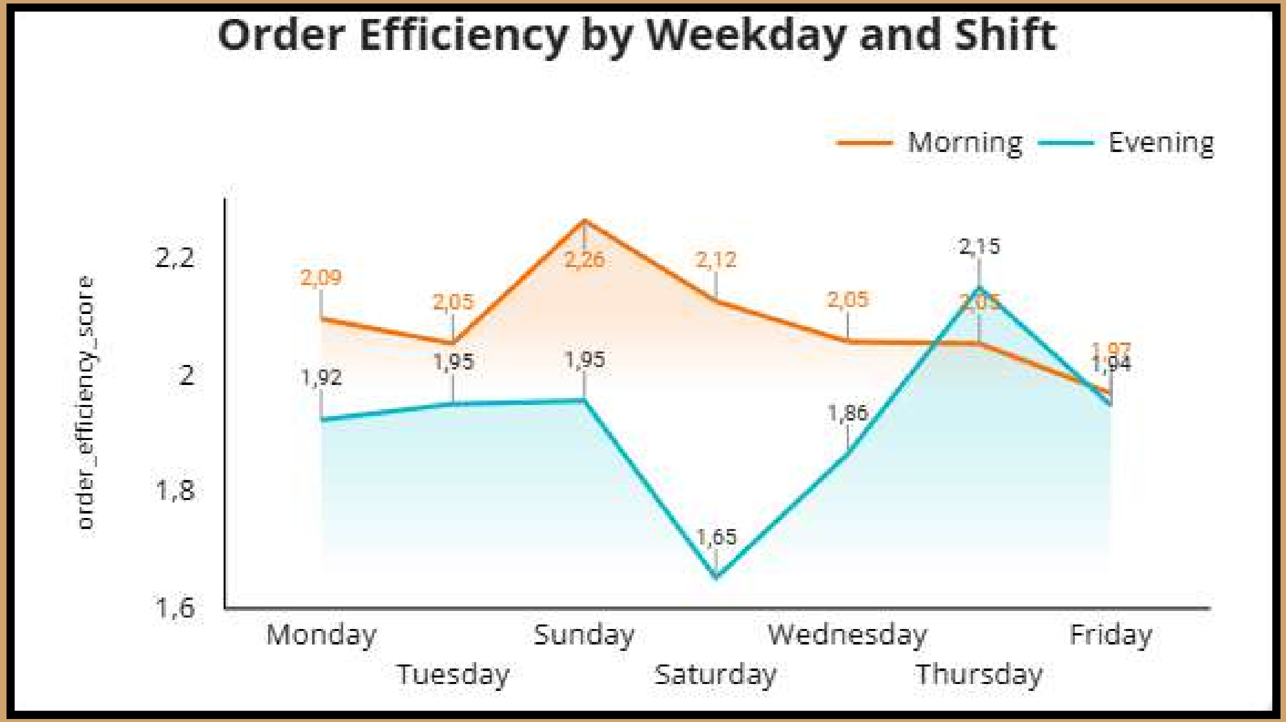
Overall Performance



overall :

- Peak performance on Sunday-Morning (2.14).
- Lowest efficiency on Friday-Evening, Saturday-Morning, Sunday-Evening.
- Weekdays (Mon-Thu) show stable efficiency (˜2.13)

Warehouse D (USA and Error Order)



Total Order and Error Percentage

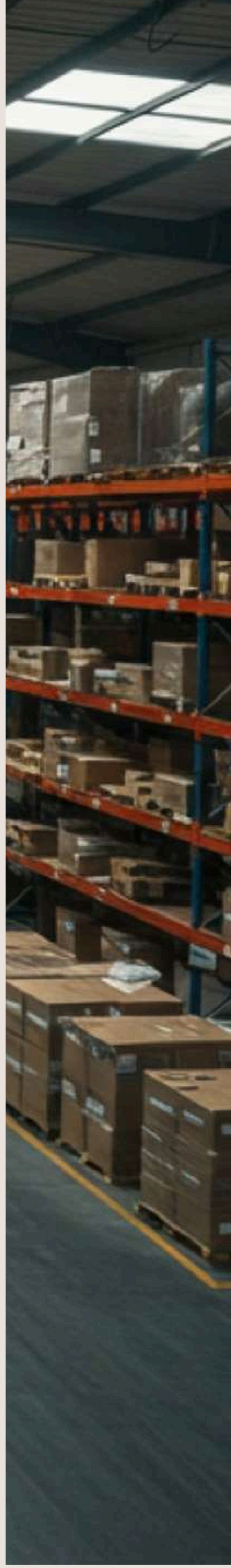
weekday_name	shift_label	total_orders	error_percentage
Monday	Evening	247	5.26
Thursday	Evening	237	5.06
Wednesday	Morning	242	4.96
Sunday	Evening	81	4.94
Monday	Morning	268	4.85
Friday	Evening	250	4.4
Saturday	Morning	126	3.97
Tuesday	Evening	256	3.91
Wednesday	Evening	220	2.73
Tuesday	Morning	264	2.65
Friday	Morning	242	1.65
Thursday	Morning	252	1.19
Saturday	Evening	87	1.15
Sunday	Morning	106	0.94

Recommendations:

- Rebalance USA orders toward morning shifts.
- Strengthen evening supervision & QC checks.
- Offer short rest breaks during evening hours.
- Provide targeted coaching for Monday/Thursday evening teams

In **Warehouse D** (USA, Error Orders) :

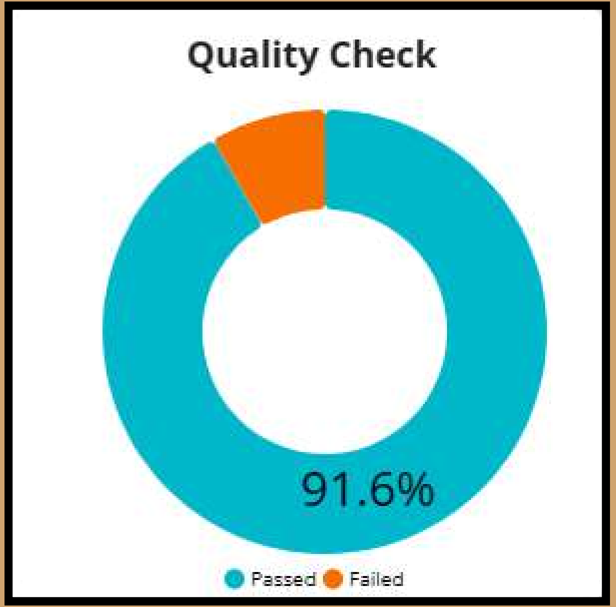
- Highest errors on Monday-Evening (5.26%) & Thursday-Evening (5.06%).
- Evening shifts = higher error rates & lower efficiency.
- Sunday-Morning shows lowest errors (0.94%) with highest efficiency (2.26).



QUALITY CHECK (QC) PERFORMANCE

Overall QC Performance

Error Rate
3,02%



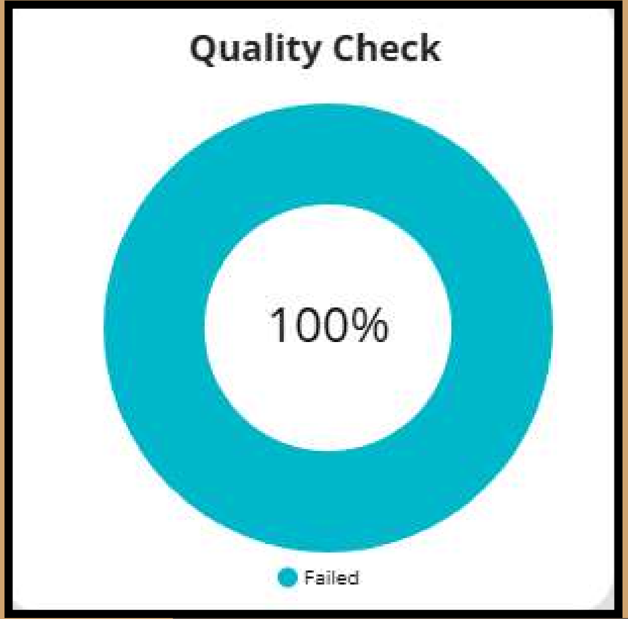
Overall

- 91.6% Passed | 8.4% Failed
- Indicates QC is generally effective — most orders meet standards.
- QC failed rate (8.4%) > error rate (3.02%) -- means QC successfully catches potential issues early.

Insight:

QC acts as a preventive filter, but failed cases suggest inconsistencies in picking or labeling

Error Order QC Performance



Warehouse D – USA (Error Orders)

- 100% of error orders failed QC → all errors were detected during inspection.

Insight:

QC is effective as a final safety net, but upstream process control is weak.

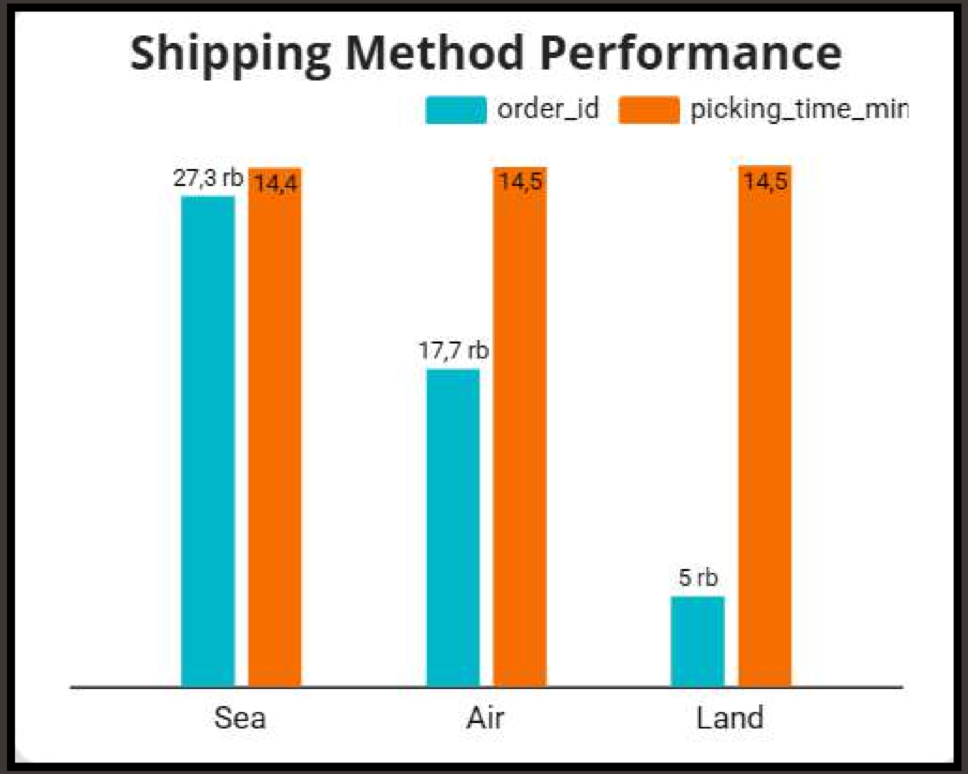
Recommendations:

- Align QC rejection criteria with actual error definitions.
- Provide refresher training for pickers & labelers.
- Strengthen preventive checks in picking & labeling stages.
- Investigate Warehouse D workflows to reduce recurring QC failures.

Since QC failures impact shipping, we analyzed how error orders are distributed across Sea, Air, and Land shipments to find where most issues occur.

SHIPPING PERFORMANCE

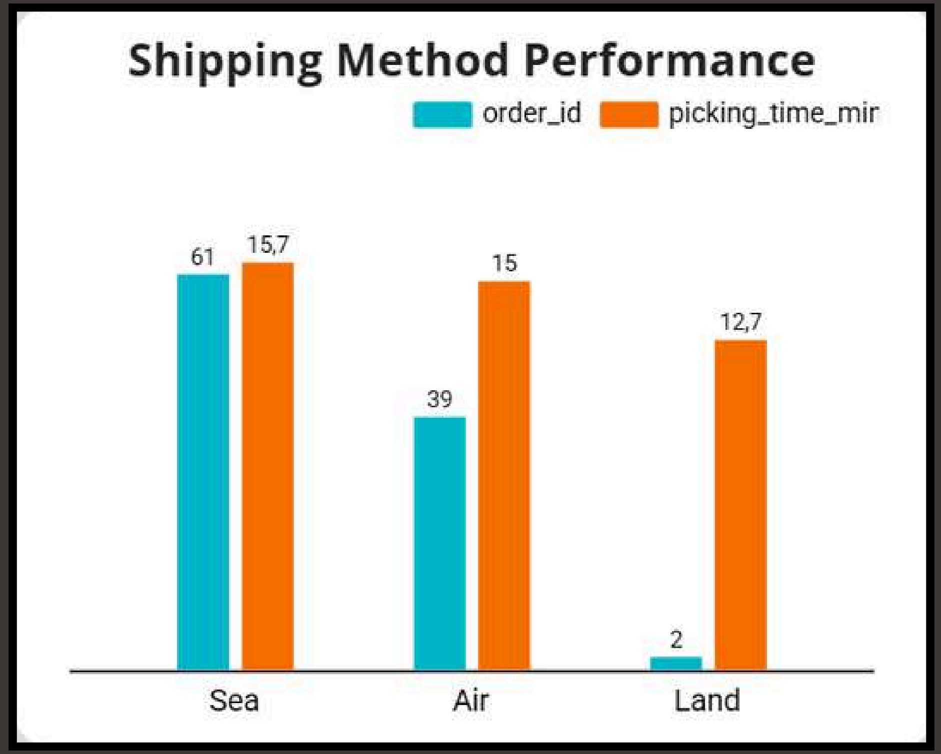
Overall Shipping Performance



Overall Performance

- Sea shipping dominates with 27.3K orders (avg. picking time: 14.4 min)
- Air: 14.5K orders (avg. picking time: 14.5 min)
- Land: 5K orders (avg. picking time: 14.5 min)

Error Order Shipping (USA and Warehouse D)



Warehouse D (USA, Error Orders):

- Sea: Error rate 3.92%, 61 error orders, avg. pick time 15.7 min
- Air: Error rate 3.74%, 39 error orders, avg. pick time 15.0 min
- Land: Error rate 0.72%, 2 error orders, avg. pick time 12.7 min

shipping_method	error_percentage
Sea	3.92
Air	3.74
Land	0.72

Insights:

- Sea shipments, though handling the most orders, are most prone to errors and longer handling times.
- Air shipments show similar efficiency but slightly fewer errors.
- Land shipments are more stable but less utilized.

Recommendations:

- Focus on improving picking and labeling accuracy, since misspick (35%) and wrong label (33%) are the leading causes of QC failures.
- Distribute volume more evenly between Sea and Air to reduce bottlenecks.
- Explore scaling Land shipping due to its consistent performance.

SUMMARY INSIGHT

Main Issue Cocentration

- The United States contributes 32% of total errors, indicating systemic issues across multiple processes rather than isolated problems.

Warehouse Performance

- A → handles most U.S. error orders but remains efficient (2.21 items/min) → potential overload risk.
- D & E → lower efficiency and higher U.S. error rates → need process improvement.

Picker Performance

- Top performers: Pickers 10, 7, 2, 4, 5 (efficient and low error).
- Low performers: Pickers 6 & 8, with higher error rates (~4-5%).

Weekday and Shift Efficiency

- Efficiency peaks on Sunday morning (2.14) but drops sharply during evening shifts (Fri-Sun).
- High error rates align with these shifts → better planning & supervision needed.

Quality Check (QC)

- Overall QC pass rate = 91.6%, but 100% of error orders failed QC, proving QC catches issues but too late in the process.

Shipping Method

- Sea shipping dominates with the most orders and highest error rate (3.92%), followed by Air (3.74%).

RECOMMENDATIONS

Warehouse Optimization

- Focus process improvement on Warehouse D, using A as a best-practice model.
- Redistribute load from A to reduce overload.
- Apply batch/zone picking for high-item orders.
- Improve item mapping to reduce walking time.

Picker Development

- Use Picker 10 as a role model/trainer for low performers.
- Provide targeted training for Pickers 6 & 8.
- Review workflow to reduce long handling times (>15 min).

Shift Management

- Rebalance complex orders toward morning shifts.
- Strengthen evening supervision & QC presence.
- Offer short rest breaks and focused coaching for Mon/Thu evening teams.

Quality Control

- Add preventive checks in picking & labeling stages to reduce QC failures.

Shipping Optimization

- Distribute order volume more evenly between Sea and Air to avoid bottlenecks and handling errors.

THANK YOU

Elin Nurulita

LinkedIn : Elin Nurulita

ellinnarulita22@gmail.com

oktober 2025