Hyundai Distribution Analytics Dashboard

A Streamlit dashboard for **real-time delivery performance analysis** with **automatic column mapping** and **robust outlier detection**. Lead time is calculated from **TOS Date → ATA Date** (with legacy fallback to **ETD → ATA** when TOS/ATA are unavailable).

🚀 Features

- **Auto column mapping** for common logistics headers (VIN/ID, Model, Destination, City/Province/Region, Transporter, TOS, dates).
- **Flexible date sources**: prefers `TOS Date` (start) and `ATA Date` (end); falls back to `ETD` and `ATA` if needed.
- **Outlier detection**:

  - **SLA-based** (per zone: Indonesia Barat / Tengah / Timur).

  - **Sigma (MAD)-based** robust statistics.

  - **Combined** mode (flags if either rule is violated).

- **Interactive analytics** across three tabs:

  1. **Overview** — KPIs, monthly trend, Top-15 Cities & Models.

  2. **Performance Analysis** — group-by (Province/Model/Transporter/City) with mean/median/std, volumes, and outlier rates.

  3. **Outlier Analysis** — filterable full records, quick downloads, and top offenders.

- **Region → Zone** auto-classification (Barat/Tengah/Timur) from `Region` text.

> Note: The former **SLA Simulation** tab is removed; SLA standards are still applied for detection when that method is selected.

📦 Installation

```
# Python 3.9+ recommended

python -m venv .venv
```

```
source .venv/bin/activate   # Windows: .venv\Scripts\activate


pip install -U streamlit pandas numpy plotly openpyxl
```

> `openpyxl` is required to read `.xlsx` files.

▶️ Run

```
streamlit run app.py
```

The app opens in your browser. Upload a CSV/XLSX and confirm column mappings in the sidebar, then click **“Analyze Data”**.

📜 Data Requirements & Mapping

Required (flexible):

- **VIN/ID**
- **Model**
- **Destination (Outlet/PDC)**
- **Start date**: `TOS Date` *(preferred)* **or** `Date ETD` *(legacy)*
- **End date**: `ATA Date` *(preferred)* **or** `Date ATA` *(legacy)*

Optional (recommended):

- `City`, `Province`, `Region` (used for grouping & zone classification)
- `Transporter`, `TOS`, `Date ATD`, `Date ETA`

Automatic detection keywords are defined in `PRIORITY`. You can rename your headers freely; the app will try to match them (and you can override via the sidebar).

Example data is embedded in the app (and downloadable as a CSV template).

📐 Calculations

Lead Time

For each row *i*:

```
lead_time_i = (ATA_i - TOS_i) in days
```

If `TOS Date` / `ATA Date` are missing, the app falls back to `ETD` / `ATA`.

Negative or missing lead times are dropped during cleaning.

Summary Stats

```
Average Lead Time = mean(lead_time)

Median  Lead Time = median(lead_time)
```

Outlier Detection

You can pick one of three modes in the sidebar:

1. **SLA-based (Custom Standards)**

Assign SLA (days) by zone (Barat/Tengah/Timur). A record is an outlier if:

```
is_outlier_i = (lead_time_i > SLA_standard_{zone_i})
```

2. **Sigma-based (Statistical, robust)** using MAD:

```
m        = median(lead_time)

MAD_raw  = median(|lead_time_i - m|)

sigmaMAD = 1.4826 × MAD_raw


is_outlier_i = |lead_time_i - m| > (k × sigmaMAD)    # default k = 3.0
```

Fallback when `sigmaMAD = 0` (near-constant data): use IQR rule

```
is_outlier_i = (lead_time_i < Q1 - 1.5×IQR) OR (lead_time_i > Q3 + 1.5×IQR)
```

3. **Combined (SLA + Sigma)**

Flags an outlier if it violates **either** SLA **or** Sigma rule. The app tags method as **Both / SLA Only / Sigma Only** for diagnostics.

🗺️ Zone Classification

`Region` strings are mapped into zones via simple keyword rules:

- **Indonesia Barat**: contains one of `sumatra/sumatera/jawa/java/bali`
- **Indonesia Tengah**: contains `kalimantan/sulawesi/nusa tenggara`
- **Indonesia Timur**: contains `papua/maluku`
- Else → `Unknown`

You can customize these rules in `classify_region_zone()`.

📊 Tabs Overview

1. **Overview**

- KPIs: total shipments, average & median lead times, outlier count & %

- Monthly trend: average & median lead time

- Top-15 Cities and Top-15 Models (by volume, color-encoded by Avg LT)

2. **Performance Analysis**

- Group by **Province / Model / Transporter / City**

- Shows **Avg/Median/Std**, **Volume**, **Outliers** and **Outlier %**

- Sorts by Avg LT (desc), top 20 groups

3. **Outlier Analysis**

- Quick KPIs for outlier subset

- Filters for **Model** (and **City**, if available)

- Full outlier records (all columns), download as CSV

- Top Models/Cities by outlier count

🛠️ Customization Tips

- **Header detection**: extend `PRIORITY` with your own synonyms.
- **Default sigma threshold**: change the `mad_threshold` slider default (3.0) in the sidebar.
- **SLA standards**: adjust defaults (e.g., 5 / 10 / 15 days).
- **Grouping choices**: edit the `group_options` section to add/remove dimensions.
- **Color scales & layout**: tweak Plotly settings for corporate themes.

❗ Troubleshooting

- **Date parsing issues**: The app first parses with `pd.to_datetime(errors="coerce")`, then retries with `dayfirst=True`. Ensure your date columns are consistent; if needed, remap columns in the sidebar.
- **All lead times are zero or identical**: MAD can be zero; the app automatically switches to IQR.
- **Negative lead time rows dropped**: Check for swapped start/end columns or incorrect dates.
- **Large files**: Consider filtering date ranges or splitting files by month to improve responsiveness.

📁 Suggested Project Structure

```
your-project/
├─ app.py                  # Streamlit app
├─ README.md               # This file
```

```
├─ requirements.txt        # Optional; pin dependencies
└─ data/                   # Optional; store sample/import files
```

requirements.txt (optional)

```
streamlit
pandas
numpy
plotly
openpyxl
```

📜 License

Choose a license (e.g., MIT) and place it here.

📝 Changelog

- **2025-10-05**: Prefer **TOS Date** → **ATA Date** for lead time; removed **SLA Simulation** tab; clarified README.