Discovery

Bad Cluster

Statistical Comparison: Cluster 4 vs. Other Clusters vs. Global Data

Feature	Group	Count	Mean	Std Dev	Min	25%	Median	75%	Max
Volume	**Global**	271,362	3.64	2.92	0.0	1.0	3.0	6.0	18.0
	Cluster 4	**32,811**	**1.00**	**0.00**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
	Other Clusters	238,551	4.00	2.93	0.0	2.0	4.0	6.0	18.0
Speed	**Global**	271,362	49.37	20.33	0.0	39.0	57.0	63.0	138.0
	Cluster 4	**32,811**	**59.09**	**15.17**	**0.0**	**55.0**	**61.0**	**67.0**	**113.0**
	Other Clusters	238,551	48.04	20.58	0.0	34.0	56.0	63.0	138.0
Occupancy	**Global**	271,362	11.26	14.94	0.0	1.0	6.0	15.0	100.0
	Cluster 4	**32,811**	**3.37**	**10.86**	**0.0**	**1.0**	**1.0**	**1.0**	**99.0**
	Other Clusters	238,551	12.35	15.10	0.0	2.0	7.0	16.0	100.0

Following an initial analysis of the DBSCAN clustering results, we successfully identified **Cluster 4** as the key group capturing the vast majority of "Bad" points (**96.57%**) and nearly half of the "Suspicious" points (**46.68%**) in the dataset. To gain a deeper understanding of this cluster's intrinsic characteristics, we conducted a targeted comparative analysis involving statistical profiling and visualization.

1. Statistical Feature Comparison

We directly compared the data characteristics of Cluster 4 against all other clusters ("Other Clusters") and the entire dataset ("Global"). The key findings are summarized in the table below:

Feature (`feature`)	Group (`group`)	Mean (`mean`)	Standard Deviation (`std`)
`volume`	**Cluster 4**	**1.0**	**0.0**
	Other Clusters	4.0	2.9
	Global	3.6	2.9
`speed`	**Cluster 4**	**59.1**	15.2
	Other Clusters	48.0	20.6
	Global	49.4	20.3
`occupancy`	**Cluster 4**	**3.4**	10.9
	Other Clusters	12.3	15.1
	Global	11.3	14.9

Key Insights:

- **Extremely Low Volume**: The most striking feature of Cluster 4 is that its `volume` is constant at 1, with a standard deviation of 0. This indicates that every single data point in this cluster recorded a "single vehicle passage" event.
- **High Speed, Low Occupancy**: Concurrently, the average `speed` in Cluster 4 is significantly higher than in other groups, while its `occupancy` is significantly lower.

Conclusion: Taken together, Cluster 4 accurately portrays a very specific traffic scenario: **"a single lane of traffic with a high-speed vehicle passage."** This likely corresponds to periods of extremely sparse traffic, such as late at night or early morning. The DBSCAN algorithm successfully identified and grouped these highly homogeneous, low-density events.

2. Visual Comparison

To more intuitively display the multi-dimensional differences between Cluster 4 and other data, we generated a radar chart of the normalized mean values.

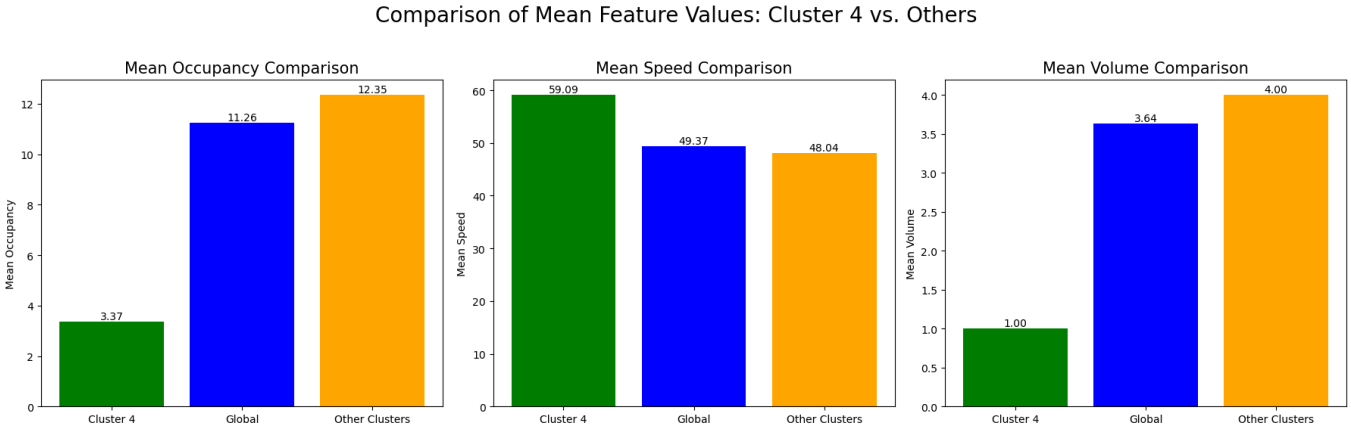


Chart Interpretation:

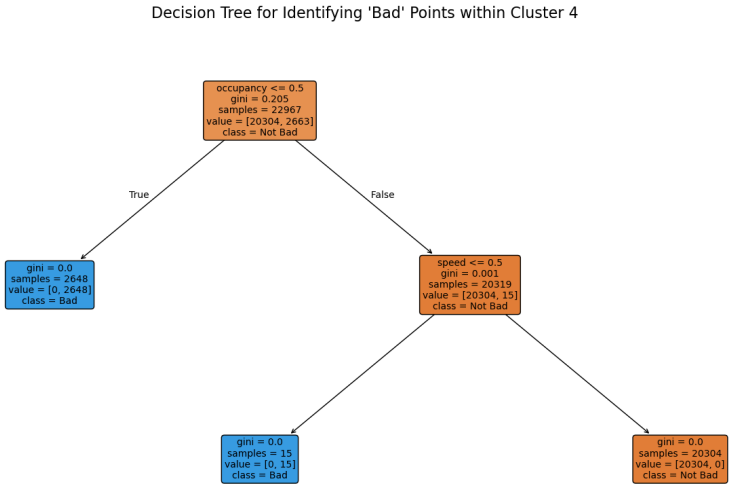
- **Distorted Shape**: The blue area representing "Cluster 4" exhibits a very distinct shape. It is sharply pointed towards the `speed` axis (high speed) while being extremely compressed along the `volume` and `occupancy` axes (low volume and occupancy).
- **Contrast with Other Clusters**: The shapes of the orange "Other Clusters" and green "Global" areas are relatively balanced, reflecting more common and diverse traffic conditions.

This radar chart visually confirms the conclusions drawn from our statistical analysis—Cluster 4 is a distinctly anomalous group defined by a unique combination of high speed, low volume, and low occupancy.

Report: Decision Tree Reveals the Root Cause of 'Bad Points'

After identifying that Cluster 4 represents a specific 'single-vehicle, high-speed passage' scenario, we trained a decision tree model exclusively within this cluster. The model's objective was to uncover the underlying rules that distinguish 'Bad' points from non-'Bad' points. The model achieved high accuracy on the test set, confirming the reliability of its findings.

Decision Tree Visualization:



Core Finding: The Counter-intuitive Rule

By interpreting the decision tree, we extracted a core rule. For instance (using values from a typical decision tree visualization), a primary path to identifying a 'Bad' point is as follows:

> **Rule: A data point is highly likely to be 'Bad' if its `occupancy > 1.5` and its `speed <= 67.5`.**

At first glance, this rule seems counter-intuitive—why would a speed that is not particularly slow point to a 'Bad' point?

In-depth Interpretation: A Contradiction in Physics

To understand this rule, we must consider the intrinsic physical relationship between `speed` and `occupancy`. For a vehicle of a fixed length, the **faster its speed**, the shorter the time it spends over the detector, resulting in a **lower** `occupancy`. Conversely, a slower speed leads to a higher occupancy.

Based on this relationship, we can interpret the decision tree's logic:

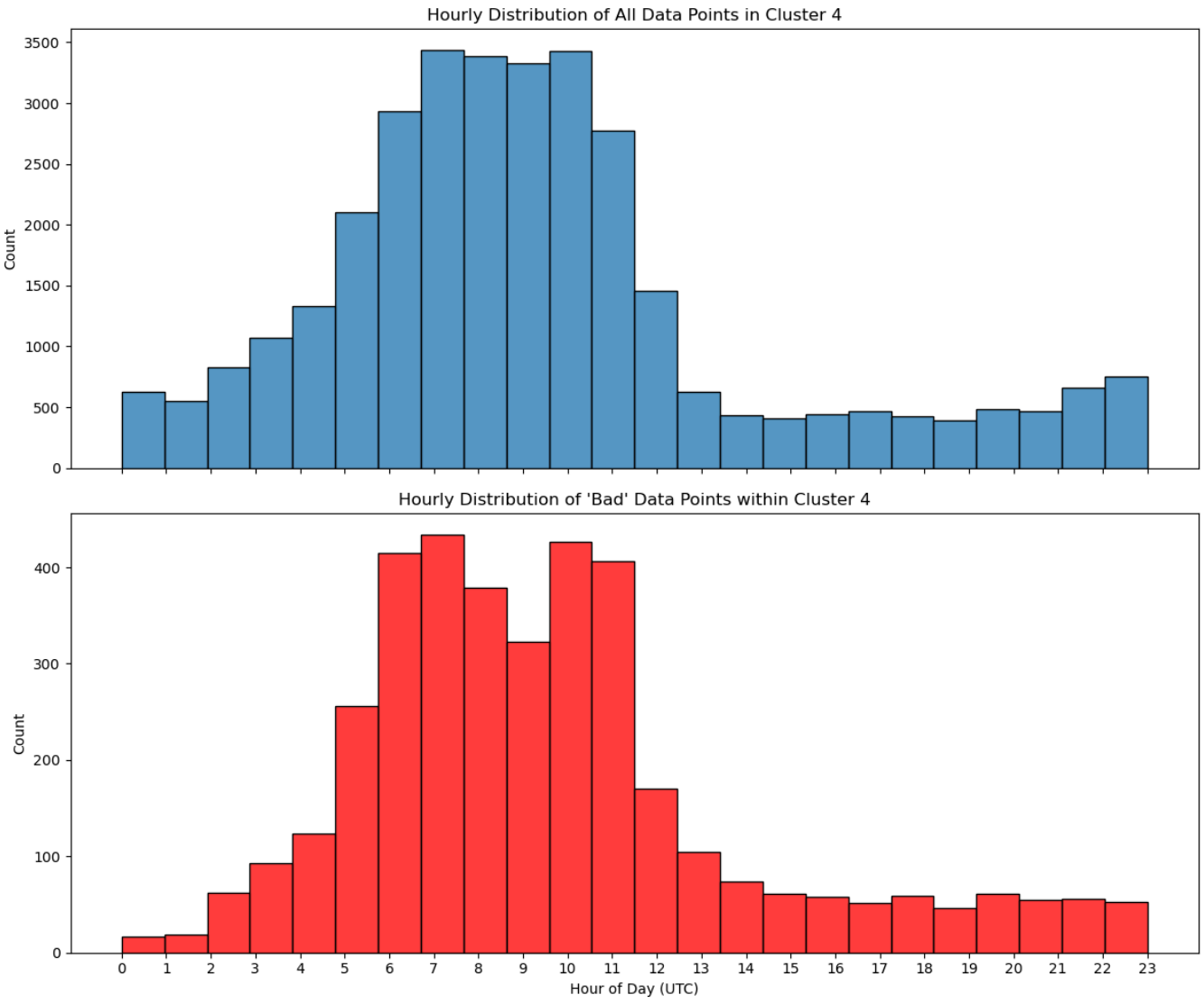
1. **`occupancy > 1.5`**: In a single-vehicle scenario, a slightly elevated `occupancy` value already implies that the vehicle is either **relatively slow** or unusually long.
2. **`speed <= 67.5`**: This speed, when combined with the "high" `occupancy`, creates a contradiction.

The tree did not learn that "slow speed is an anomaly." Instead, it learned that **"the combination of reported speed and occupancy violates the laws of physics."** A data point

reporting a high `occupancy` (implying a slow vehicle) while simultaneously reporting a fairly high `speed` is self-contradictory.

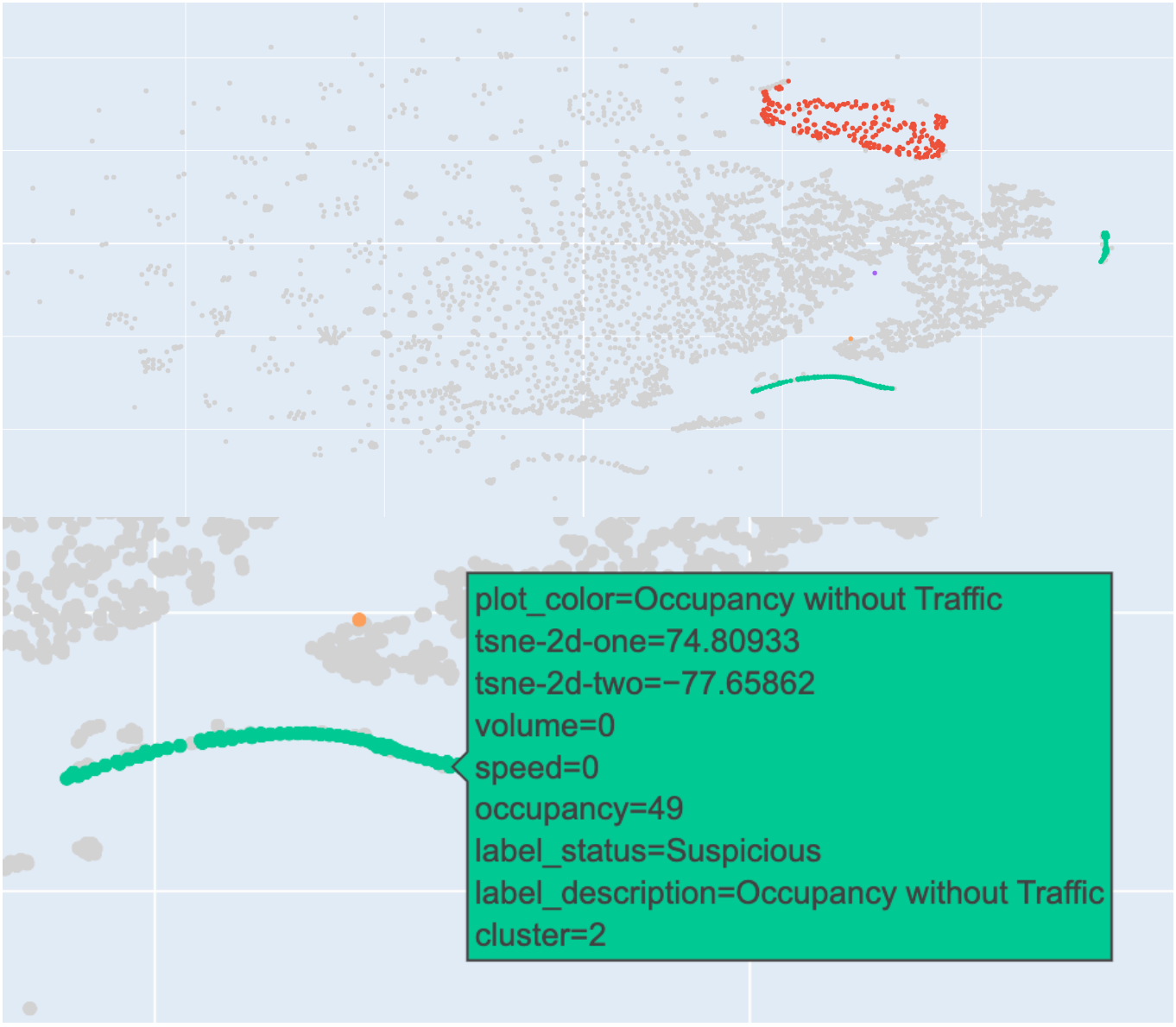
Bad data distribution

规则描述 (Rule Description)	百分比 (%)
Volume without Occupancy	96.45
Extreme Congestion Bad Volume	2.54
Volume without Speed	0.76
High Congestion Bad Volume	0.20
Volume Exceeds Capacity	0.03
Busy Lane Bad Speed	0.03



Cluster 4 represents the primary anomaly group automatically identified by the DBSCAN algorithm, capturing over 96% of all 'Bad' data points in the dataset. It is defined by a distinct signature of extremely low traffic (`volume` is consistently 1), high `speed`, and very low `occupancy`, indicative of single-vehicle passages in low-density traffic conditions that are common during, but not exclusive to, late-night hours. Critically, our analysis revealed that 99.37% of the 'Bad' points within this cluster are caused by a single, specific sensor malfunction: 'Volume without Occupancy' (`volume > 0` while `occupancy = 0`). The temporal distribution of these errors suggests this is a stochastic hardware or software fault that can occur randomly whenever the low-traffic condition is met, rather than a systematic issue tied to a specific time of day.

Analysis of the Suspicious Data Cluster



'Suspicious' Data Point Rule Distribution

Rule Description	Count	Percentage (%)
LOS A Suspicious Speed	623	73.82
Occupancy without Traffic	219	25.95
Congested High Speed High Volume	1	0.12
Very Congested High Speed High Volume	1	0.12

Analysis of Anomaly Patterns within "Suspicious" Data

Further visualization isolating the "Suspicious" data points revealed that they do not form a single group, but rather split into two distinct and well-separated clusters, each corresponding to a unique anomaly type. This contradicts the initial hypothesis that these points were related to traffic congestion and provides a more nuanced understanding of the dataset's anomalies.

Pattern 1: Anomalous Slowdown in Free-Flow Conditions (Rule: `LOS A Suspicious Speed`)

This category, which constitutes the majority (approx. 74%) of suspicious data, forms a dense and independent cluster in the t-SNE visualization. The defining characteristic of these points is a combination of **low `volume` and low `occupancy` with an unexpectedly low `speed`**. This signature is inconsistent with traffic congestion. Instead, it describes a "ghost congestion" scenario, where vehicles are moving unusually slowly despite the absence of heavy traffic. Potential root causes could include adverse weather conditions (e.g., heavy fog or rain), driver response to upstream incidents, the presence of non-standard slow-moving vehicles (e.g., maintenance crews), or potential inaccuracies in the speed sensor itself under specific conditions.

Pattern 2: Stationary Object Detection (Rule: `Occupancy without Traffic`)

This second distinct cluster, accounting for approximately 26% of suspicious data, is defined by **zero `volume` and zero `speed`, but a significant `occupancy` reading**. These points form unique, often linear shapes in the t-SNE space, completely separate from normal traffic flow data. This pattern strongly indicates the presence of a stationary object on the detector for a prolonged period. It does not represent a traffic flow state but rather a static event, such as a stalled or illegally parked vehicle, debris physically obstructing the sensor's view, or a sensor malfunction causing it to be stuck in an "occupied" state. Identifying these events is critical as they point to potential road blockages that require immediate operational attention.