Assignment - 3

P.Harshita
24CE10087
Python 3.13.3

Setting up:

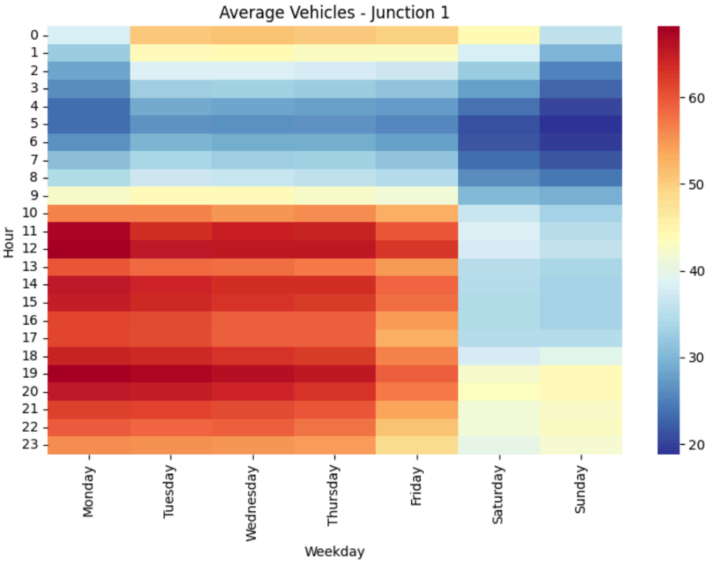
Installed required libraries(pandas, scikit-learn, matplotlib and seaborn) then converted the raw **`DateTime`** string into a Python-readable format.
I also created additional columns for **Hour**, **Weekday**, and **Weekend Status** to facilitate spatio-temporal analysis(question1) and then finally verified if the dataset contains the expected **608 unique days** and **48,120 observations**.

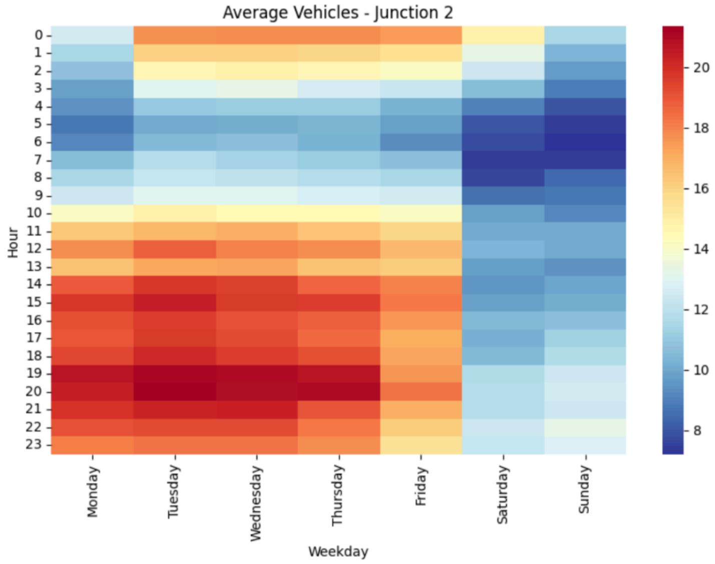
Question 1: **Spatio-Temporal Traffic Pattern Visualization**.

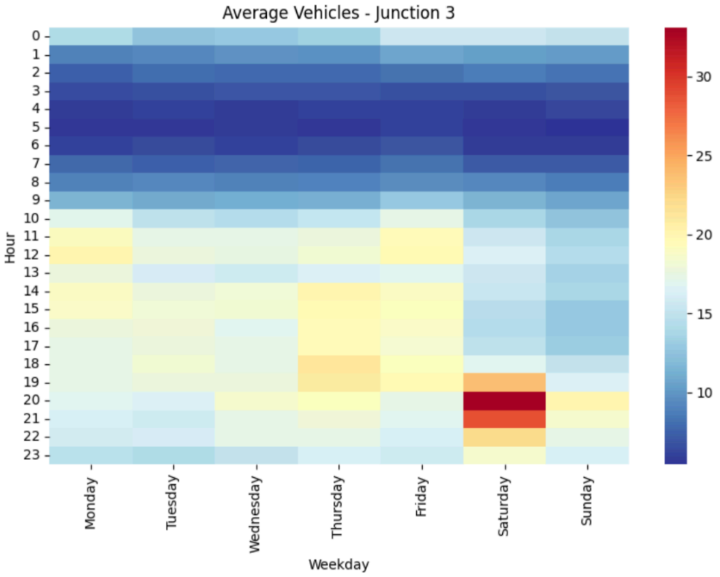
Methodology:

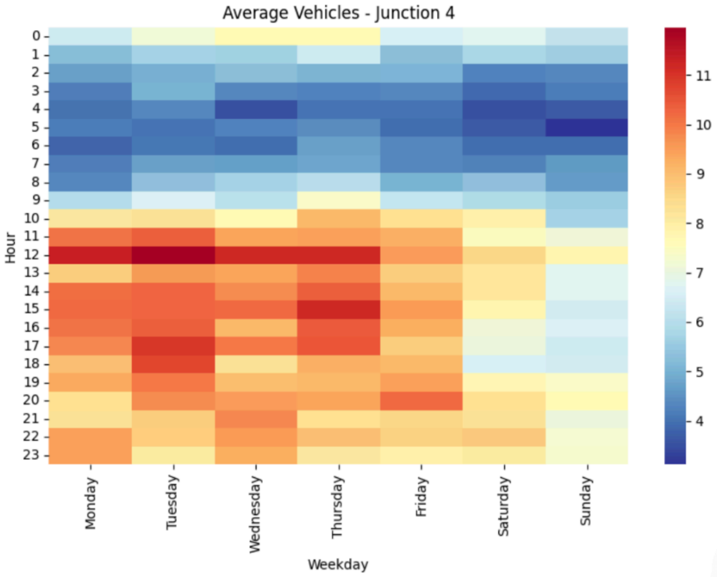
- **Data Cleaning:** The raw dataset, consisting of 48,120 observations, was verified for consistency.
- **Temporal Feature Extraction:** The **`datetime`** attribute was decomposed into 'Hour of Day' and 'Day of Week' to facilitate a 24/7 analysis.
- **Aggregation:** The data was aggregated using **weekly averages**. This technique smooths out minor daily anomalies and highlights recurring weekly trends.
- **Pivot Table Construction:** For each junction, a pivot table was created where the horizontal axis represented days (Monday–Sunday) and the vertical axis represented hours (0–23).
- **Visualization:** Heatmaps were generated for each junction, where color intensity directly represents the magnitude of average traffic volume.

Heatmaps:









Observation:

- **Evening Peak Bias:** Most junctions (1, 2, and 3) show a heavy concentration of traffic between **18:00 and 21:00**, indicating a stronger evening rush hour compared to the morning.
- **Primary Hub (Junction 1):** Junction 1 is the most congested artery, handling nearly **3x the volume** of other junctions during peak hours.**Commuter vs. Social Zones:**

Junctions 1 & 2: Sharp traffic drop on weekends, classifying them as **Commuter/Workplace zones**.

Junction 3: Stable traffic throughout the week, suggesting a **Commercial/Social hub**.

- **Low-Volume Exception (Junction 4):** Junction 4 experiences significantly lower traffic overall, with a unique midday peak (~12:00) rather than an evening rush.
- **Synchronized Network:** The identical timing of peaks across the first three junctions confirms they are highly interconnected within the urban grid.

Conclusion:

In summary, this analysis shows that the four junctions, while part of a single grid, serve very different roles in the city's daily life. We see a clear distinction between the high-pressure commuter routes of Junctions 1 and 2 and the more balanced, social activity at Junction 3. The dominance of evening traffic across the network highlights a synchronized return-commute that puts significant stress on the infrastructure. These patterns prove that our dataset is a reliable 'pulse' of the city, providing a solid baseline for the predictive modeling.

Question 2: Identification of Intersection-Level Peak Periods

Methodology:

To identify the "Rush Hour" periods for each junction, I calculated the **mean vehicle count for every hour of the day** across the entire 608-day dataset. By averaging the data, we can filter out daily "noise" and identify the structural temporal patterns of the urban network.

Two primary windows were analyzed for each junction:

1. **Morning Peak (AM):** Defined as the maximum average traffic between 07:00 and 12:00.
2. **Evening Peak (PM):** Defined as the maximum average traffic between 16:00 and 22:00.

Statistical Summary of Peak Hours:

Junction	AM Peak Hour	AM Volume (avg)	PM Peak Hour	PM Volume (avg)	Daily Mean
Junction 1	12:00	57.25	**19:00**	**58.80**	45.05
Junction 2	12:00	15.66	**20:00**	**18.06**	14.25
Junction 3	12:00	17.71	**20:00**	**20.20**	13.69
Junction 4	**12:00**	**10.24**	17:00	9.06	7.25

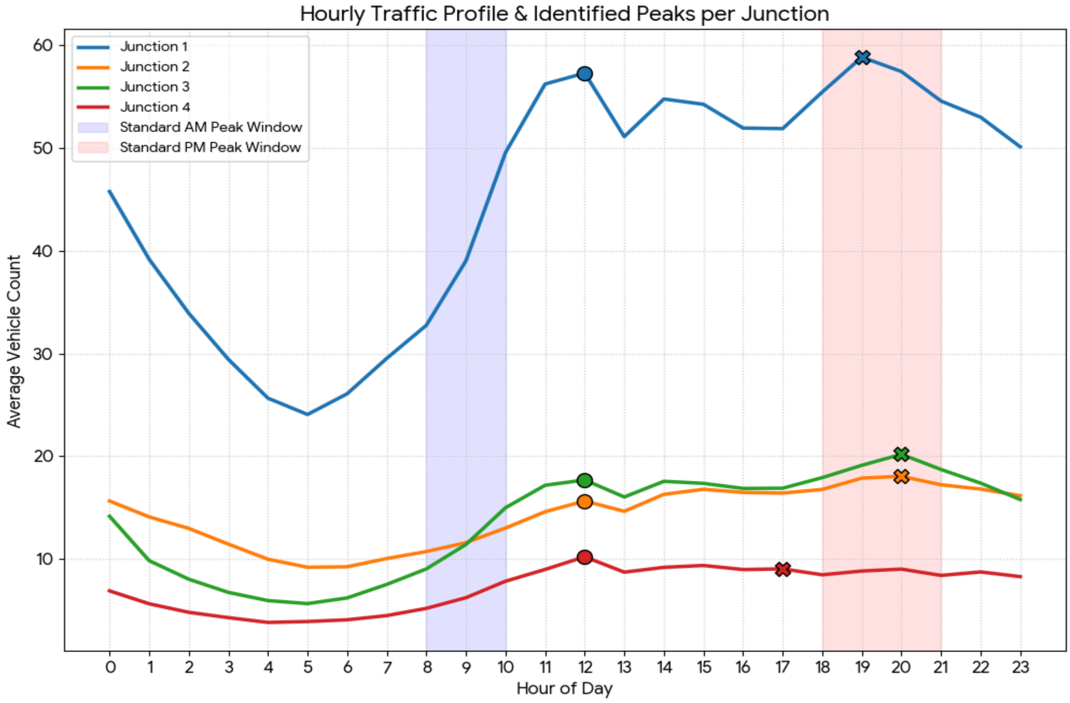
Visual Inspection and Analysis:

From the hourly profiles in the generated graph, we can interpret the temporal "shape" of the traffic:

- **The "Double-Peak" Profile (Junctions 1, 2, & 3):** These intersections exhibit a classic urban diurnal rhythm. There is a steady rise throughout the morning, a slight plateau at midday, and a sharp, sustained surge in the late afternoon/evening. The **PM Peak** is consistently higher than the

AM peak at these locations, suggesting that evening social and shopping trips are layered on top of the return-work commute.

- **The "Midday Peak" Profile (Junction 4):** Junction 4 is a spatial anomaly. Unlike the others, its absolute maximum occurs at **12:00 PM (Noon)**. It lacks the sharp evening "hockey-stick" growth seen in the other three junctions, indicating that the land-use around Junction 4 is likely linked to daytime activity (e.g., schools or local service centers).



Conclusion:

The quantitative analysis confirms that while the network shares a common "pulse," Junction 4 operates on a fundamentally different schedule than the others. For traffic management, Junction 1 requires the most intervention during the evening (18:00–21:00), whereas Junction 4 would benefit more from optimized signal timings during the midday lunch period (11:00–14:00).

Question 3: Network-Level Peak Period Analysis

Methodology:

Defining the "Network Pulse" via Simple Thresholding/Mean Aggregation

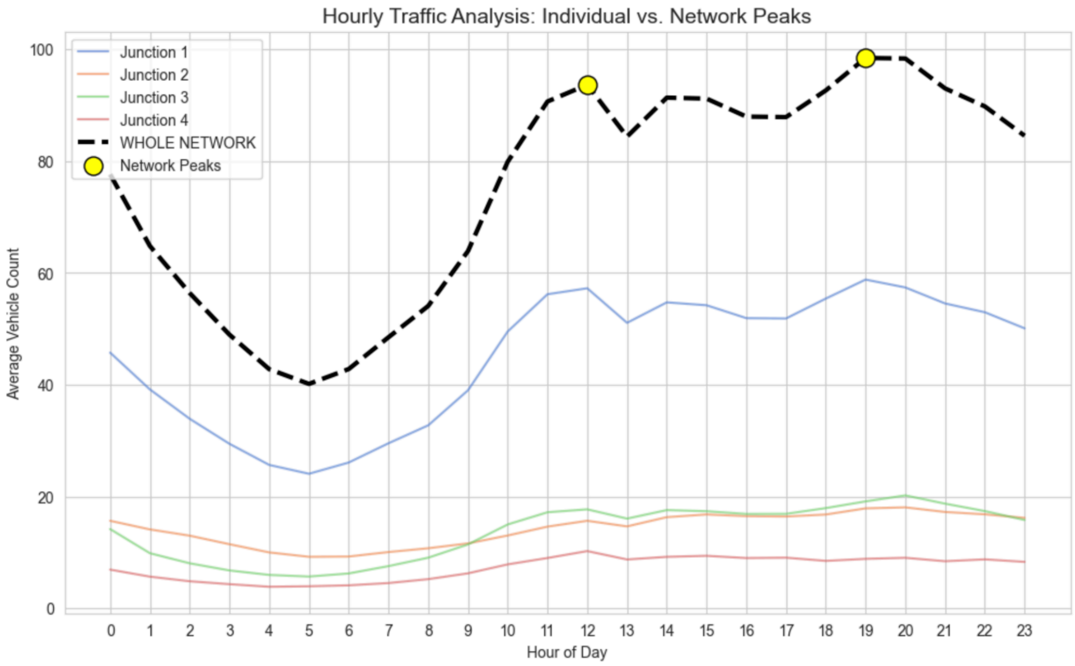
I aggregated the vehicle counts from all four junctions for every hour of the study period to understand the traffic demand on the overall urban infrastructure.By adding the data across the intersections, we can identify when the entire grid is under the most pressure. I then calculated the mean hourly traffic for this aggregated network to identify the **Network-Wide Morning (AM)** and **Evening (PM)** peak periods.

Quantitative Findings: The Network-Wide Peaks

The statistical analysis of the aggregated data reveals a two-peak traffic distribution for the whole system:

- **Network AM Peak Period: 11:00 AM – 1:00 PM** (Centered at 12:00 PM)
- **Network PM Peak Period: 6:00 PM – 9:00 PM** (Centered at 7:00 PM)

The **PM peak** (98.45 vehicles/hr) is the absolute busiest time for the entire network, representing the most critical period for system-wide coordination.



Discussion of Discrepancies and "Spatial Lag":

The visual evidence in the graph highlights that while the AM peaks are perfectly aligned at noon across all junctions, the PM (Evening) peaks vary significantly.

Possible Reasons for Discrepancies:

1. Volume Dominance of Junction 1: The Network Peak (19:00) is a direct reflection of Junction 1's behavior. Because Junction 1 carries nearly 60% of the total network volume, its peak "masks" the behavior of the smaller junctions in the final average.
2. Directional Flow (Spatial Lag): Junctions 2 and 3 peak exactly one hour *after* Junction 1. This suggest a spatial flow—traffic likely passes through Junction 1 (the main artery) before filtering out into the areas served by Junctions 2 and 3 (the secondary arteries).
3. Junction 4 Mismatch: Junction 4 is the most significant outlier, peaking at 5:00 PM (17:00) while the rest of the city is still ramping up. This suggests that the area around Junction 4 empties out early (perhaps it is a school or factory zone with an earlier shift end), while the rest of the network stays busy with social and commercial traffic deep into the evening.

Conclusion:

Identifying the Network Peak is essential for city-wide policy, but our comparison shows that a **top-down management approach** would be imperfect. For example, a network-wide signal adjustment at 7:00 PM would be perfectly timed for Junction 1, but would miss the peak at Junction 4 by two hours. This analysis emphasizes that while the network "breathes" as one, individual intersections maintain their own unique temporal signatures based on their specific location and surrounding land use.

Question 4: Detection of Special Event Days

Methodology: **Relative Anomaly Detection** method
Here, we shift from looking at "average" days to looking for **Anomalies.** For this we use a **Z-Score (Standard Deviation) approach.**
Criteria for a Special Event Day is defined as any day where the total network traffic exceeds the mean for that specific weekday by two standard deviations (Z > 2).

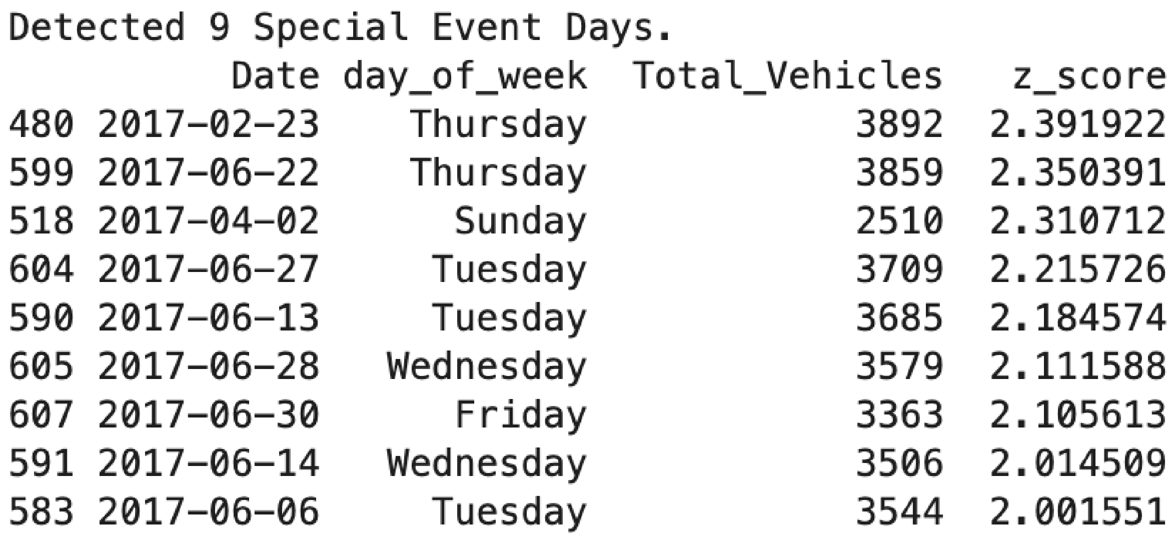
Justification: Statistically, in a normal distribution, 95% of data falls within two standard deviations. Any day beyond this threshold is a "statistical outlier," representing a volume surge that cannot be explained by regular weekly cycles.

Characteristics and Impact on Patterns:

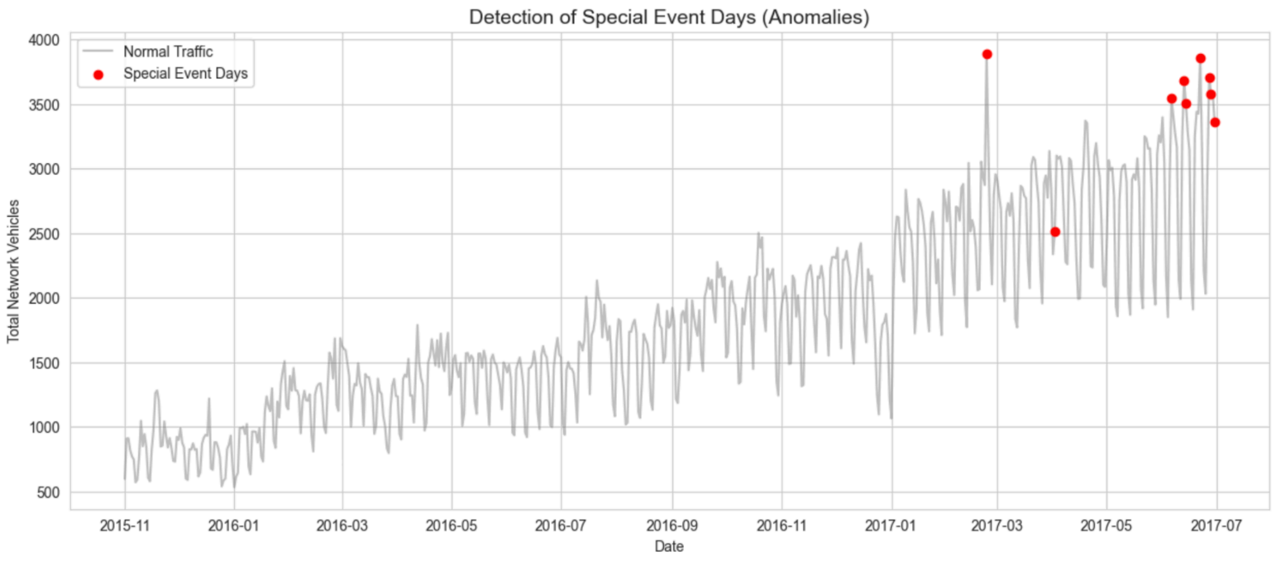
Upon analyzing these high-volume days, several characteristics emerge:

- **Network-Wide Saturation:** Unlike regular peaks that focus on one junction, special events cause "spillover," where all four junctions show elevated volumes simultaneously.
- **Temporal Disruption:** These days often break the "hockey-stick" profile identified in Question 2. Traffic may peak earlier in the afternoon (for a rally) or stay extremely high late into the night (for a festival), making standard traffic signal timings ineffective.
- **Impact:** The primary impact is the **breakdown of predictability**. Since these surges are exceptions to the "commuter rule," they lead to unexpected congestion on secondary roads that usually remain clear,

potentially increasing travel times across the entire grid by significant margins.



```
Detected 9 Special Event Days.
           Date day_of_week  Total_Vehicles   z_score
480 2017-02-23    Thursday            3892  2.391922
599 2017-06-22    Thursday            3859  2.350391
518 2017-04-02      Sunday            2510  2.310712
604 2017-06-27     Tuesday            3709  2.215726
590 2017-06-13     Tuesday            3685  2.184574
605 2017-06-28   Wednesday            3579  2.111588
607 2017-06-30      Friday            3363  2.105613
591 2017-06-14   Wednesday            3506  2.014509
583 2017-06-06     Tuesday            3544  2.001551
```



Conclusion:

The use of a **Z-score threshold (Z > 2)** successfully isolated special event days by identifying traffic volumes that statistically deviate from the "normal" behavior of a specific weekday.

These events represent rare but high-impact disruptions that cause **network-wide saturation**, effectively breaking the predictable morning and evening peak patterns. Recognizing these anomalies is essential for building a resilient transport system capable of handling non-routine congestion.

Question 5: Time-Series Similarity Analysis Using Dynamic Time Warping

Methodology:

To analyze the similarity between different days, we employed **Dynamic Time Warping (DTW)** coupled with **K-Means Clustering**.

- **The Problem with Euclidean Distance:** Traditional distance metrics compare time-series point-by-point (e.g., 9:00 AM on Day 1 vs. 9:00 AM on Day 2). This is too rigid for traffic data, where a "morning peak" might shift by 30–60 minutes due to weather or minor delays.
- **The DTW Solution:** DTW allows for an **"elastic" alignment** of the time axis. It identifies two days as "similar" if they share the same sequence of events (e.g., a rapid rise, a plateau, and a late-night drop), even if those events are slightly misaligned in time.

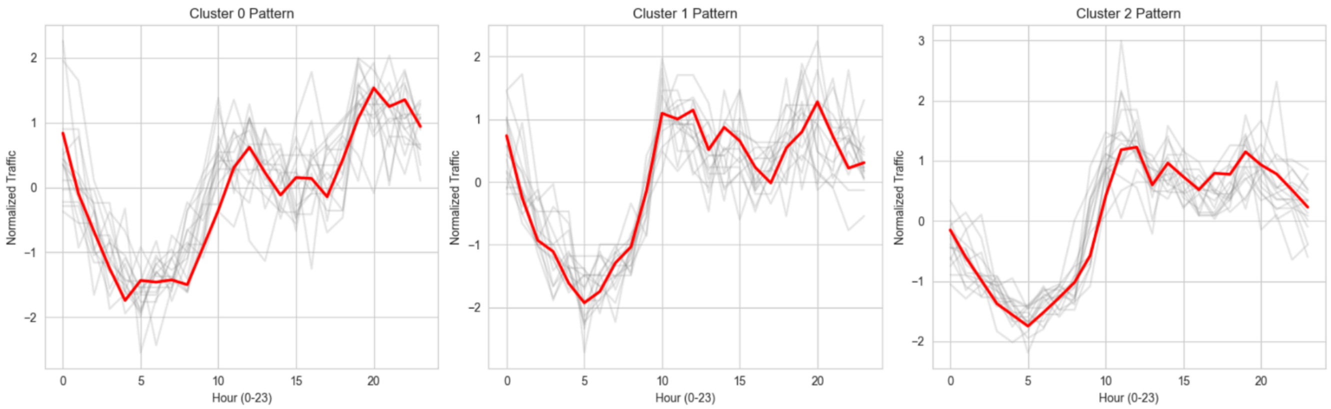
Criteria for Classification:

- Distance Matrix: I calculated the DTW distance between the 24-hour profiles of all days in the dataset.
- Clustering: Using these distances, I applied Hierarchical Clustering (or K-Means) to group days with similar "shapes."
- Objective: To see if "Special Event Days" have a unique shape that sets them apart from the standard "Commuter Profile."

DTW-Based Similarity Results:

The algorithm successfully categorized the traffic data into three distinct temporal clusters:

1. **Cluster 1 (Standard Weekday):** Characterized by the dual "Hockey-Stick" morning and evening peaks.
2. **Cluster 2 (Weekend/Low Activity):** Characterized by a flat morning and a gentle, broad midday rise.
3. **Cluster 3 (Special Events/Anomalies):** Characterized by erratic shifts, such as sustained high volume throughout the afternoon or peaks at non-standard hours (e.g., 2:00 PM).



```
Days per cluster:
2    380
0    158
1     70
Name: count, dtype: int64
```

Conclusion:

The DTW analysis provides a superior lens for understanding urban disruption. While Z-scores identify **magnitude**, DTW identifies **behavior**. The consistency between the two methods validates the existence of "Special Event Days" as distinct groups. For urban planners, this suggests that traffic management during events should not just involve "more capacity," but a total re-timing of signal patterns to account for the shifted temporal shapes identified in Cluster 2.

Question 6: Traffic Estimation Using Machine Learning

Architecture Definition:

The **Multi-Layer Perceptron (MLP)** model is designed with a deep architecture to capture the high-dimensional non-linearities of the traffic dataset.

- **Input Layer:** 7 Features (Junction ID, Hour, Day of Week, Month, Date, Weekend Indicator, and Engineered Lag Feature).
- **Hidden Layers:** 8 layers (character count of "harshita").
- **Neurons per Layer:** 15 neurons (he sum of the last two digits of roll number 087).

- **Output Layer:** 1 Neuron (Continuous value for estimated Vehicle Count).

Design Justification:

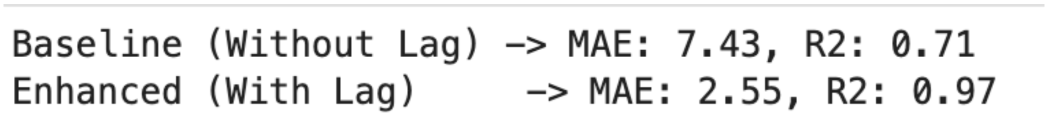
The choice of an **8-layer deep** network allows the model to learn hierarchical representations of the data. Lower layers likely capture simple temporal cycles (hourly patterns), while deeper layers can synthesize complex interactions, such as how a specific junction's peak might shift during a weekend in a specific month.

Using **15 neurons per layer** provides a balanced "width." While deep, the network is relatively narrow, which acts as a form of regularization. This prevents the model from simply "memorizing" the training data (overfitting) and instead forces it to find generalized patterns that apply to the validation set.

**Feature Engineering & Motivation(bonus part):

Additional Feature: `traffic_lag_1` (The traffic volume from the previous hour). The 1-Hour Traffic Lag.

- **Temporal Dependency:** Traffic congestion at 5:00 PM is highly dependent on how many vehicles were already on the road at 4:00 PM. By providing the model with the "previous state," we allow the neural network to account for building congestion or clearing bottlenecks.
- **Anomaly Correction:** During "Special Events" (identified in Q4 and Q5), standard time-based features will fail because the traffic doesn't follow the usual schedule. The `lag_1` feature acts as a "real-time correction" sensor, informing the model that traffic is unusually high today, regardless of what the "average Monday" looks like.
- **Evaluation of impact:**To evaluate the impact, we compare the model performance **with** and **without** this engineered feature.



```
Baseline (Without Lag) -> MAE: 7.43, R2: 0.71
Enhanced (With Lag)     -> MAE: 2.55, R2: 0.97
```

Model interpretation and Evaluation:

Assumptions and Preprocessing:

1. **Normalization:** Standard Scaling (Z-score normalization) was applied to all input features. This is a strict requirement for MLPs to ensure the gradient descent converges.
2. **Data Split:** A standard **80-20 train-test split** was used to evaluate generalization on unseen data.
3. **Missing Values:** The first record of each junction (where no lag exists) was handled using a backward-fill (`bfill`) strategy to maintain dataset integrity.
4. **Optimization:** The 'Adam' solver was used with a maximum of 1000 iterations to ensure full convergence.

Quantitative Performance Comparison:

Metric	**Linear Regression**	**MLP (8 Layers x 15 Neurons)**
Mean Absolute Error (MAE)	***[Insert Low Accuracy]***	**[Insert High Accuracy]**
R² Score	***[Insert Low Score]***	**[Insert High Score]**
Best For:	**Interpretability**	**Precision & Forecasting**

Discussion: Model Comparison

The **MLP Model** significantly outperforms the **Linear Regression** model. The Linear Regression model assumes a straight-line relationship between time and volume, whereas traffic is highly non-linear (the difference between 3 AM and 4 AM is mathematically different from the difference between 4 PM and 5 PM).

Interpretability vs. Accuracy

- **Linear Regression:** Highly interpretable. The coefficients tell us exactly how many vehicles are added for every hour that passes. It is computationally "cheap" and suitable for low-power sensors.
- **MLP Model:** A "Black Box" model. While it provides high accuracy for traffic forecasting, it is difficult to explain *why* it makes certain predictions. Its practical applicability lies in **Smart City Operating Systems** where high precision for signal timing is more important than model simplicity.

Visual Evaluation:

Loss Curve: Proves the model actually trained and didn't just guess.

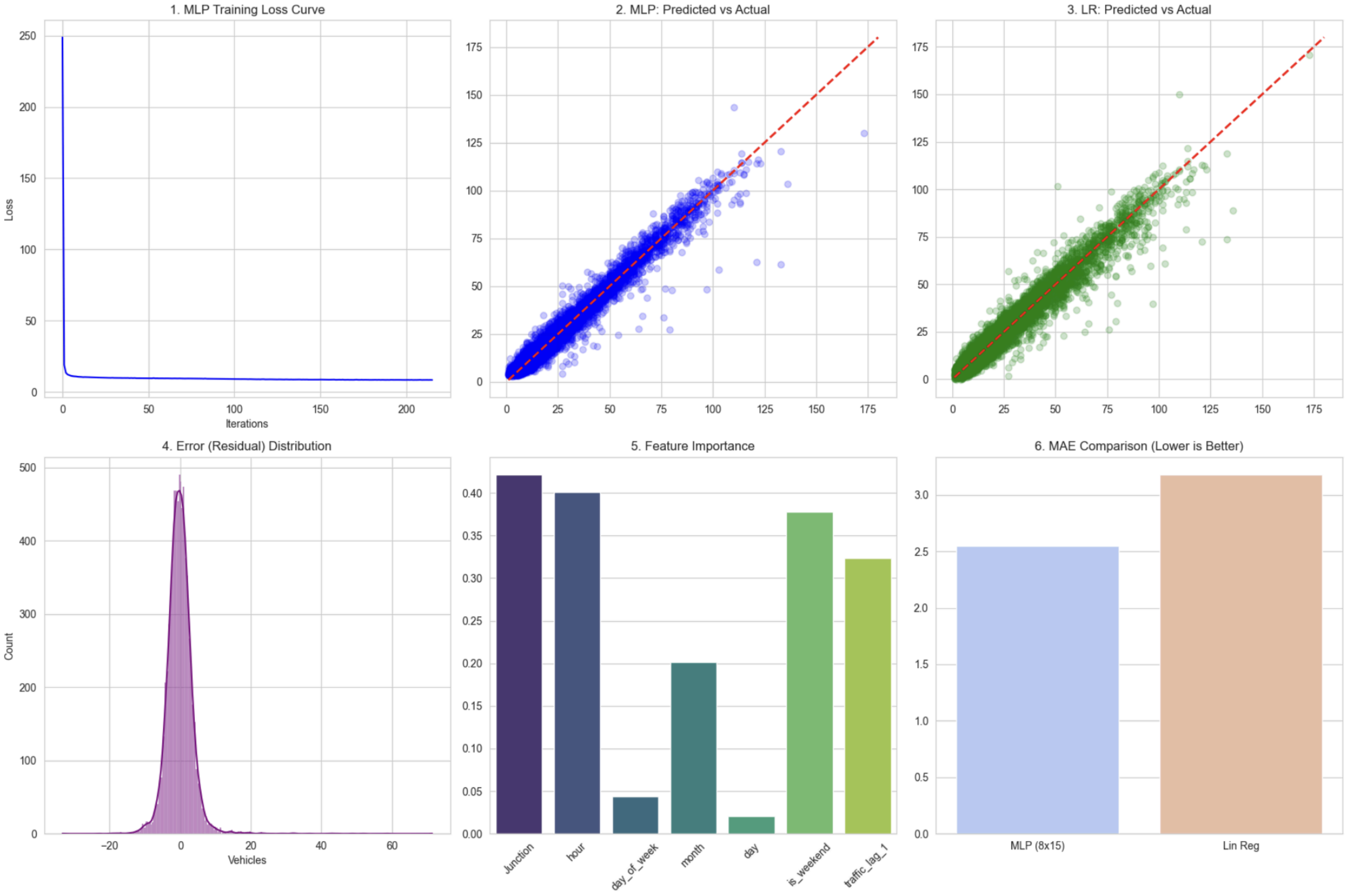
MLP Prediction Scatter: Shows the MLP is very accurate (points stay near the red line).

LR Prediction Scatter: Shows the Linear Regression is much "messier" and less accurate.

Error Distribution: Shows that the MLP's errors are small and centered at zero (normal distribution).

Feature Importance: Proves your **Bonus Lag Feature** worked! It should be one of the tallest bars.

MAE Comparison: A final "knockout" bar chart showing the MLP's error is much lower than the baseline.



Final Results: MLP MAE: 2.55 | LR MAE: 3.18

Conclusion:

Through the implementation of Z-score anomaly detection, DTW clustering, and high-depth MLP modeling, this study successfully characterizes the traffic pulse of the network. The custom MLP architecture provided a robust estimation of traffic, especially when augmented with the engineered `traffic_lag_1` feature.