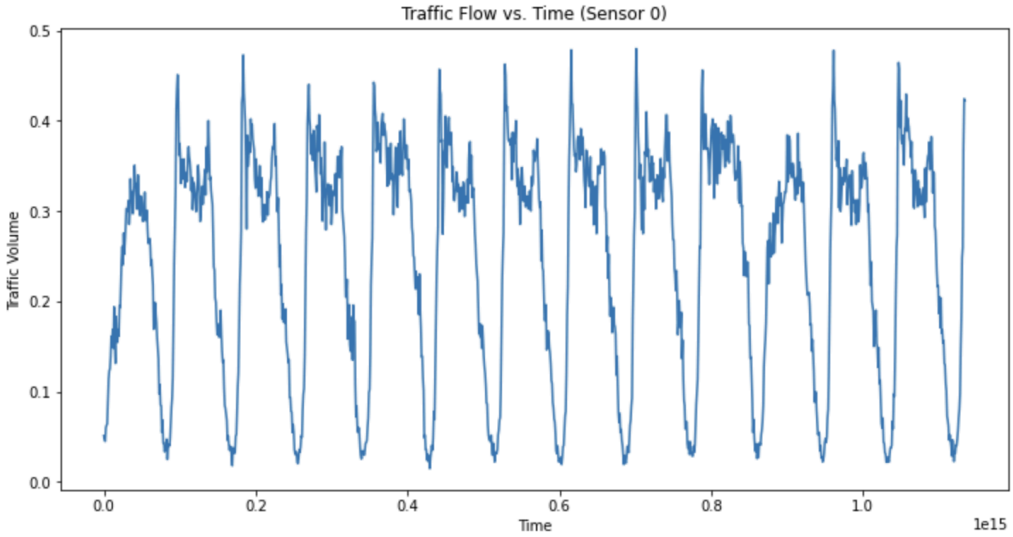
3. Exploratory Data Analysis

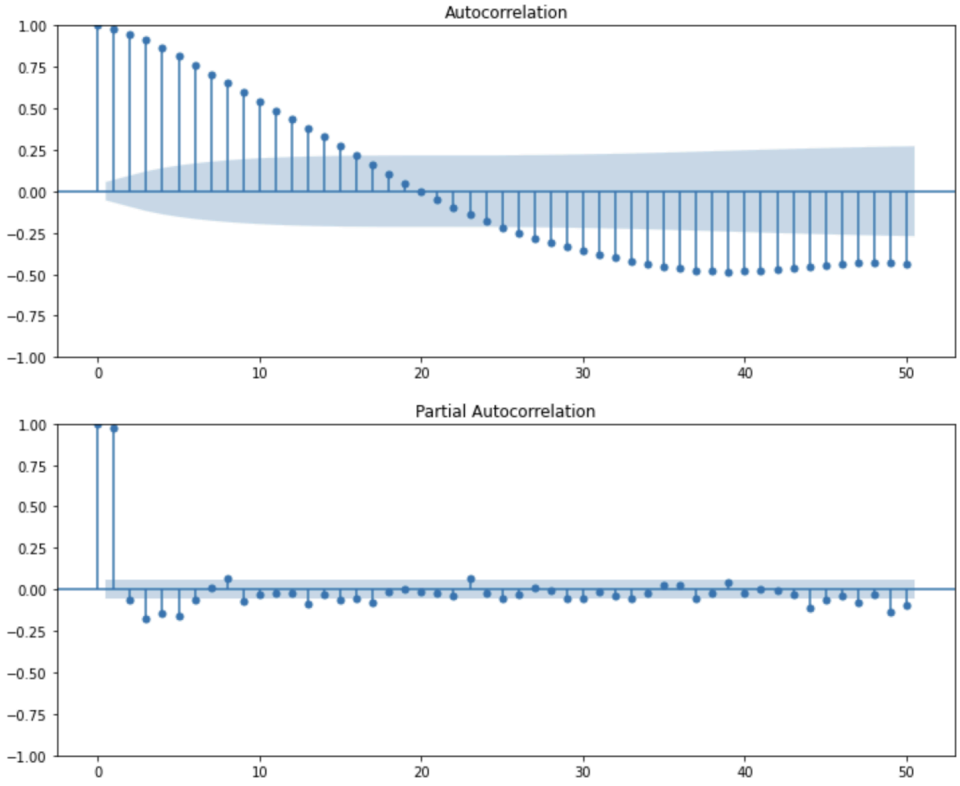
3.1 Time Series Analysis

- **Traffic Flow vs. Time (Sensor 0)**: The time series plot for Sensor 0 shows clear cyclical patterns in traffic flow, indicating regular fluctuations over time. There are periods of high traffic followed by sharp drops, which suggests that traffic volume follows a predictable daily or weekly cycle. Peaks likely represent times of high traffic demand (e.g., rush hours), and troughs may correspond to periods of lower traffic, such as late at night or early in the morning.



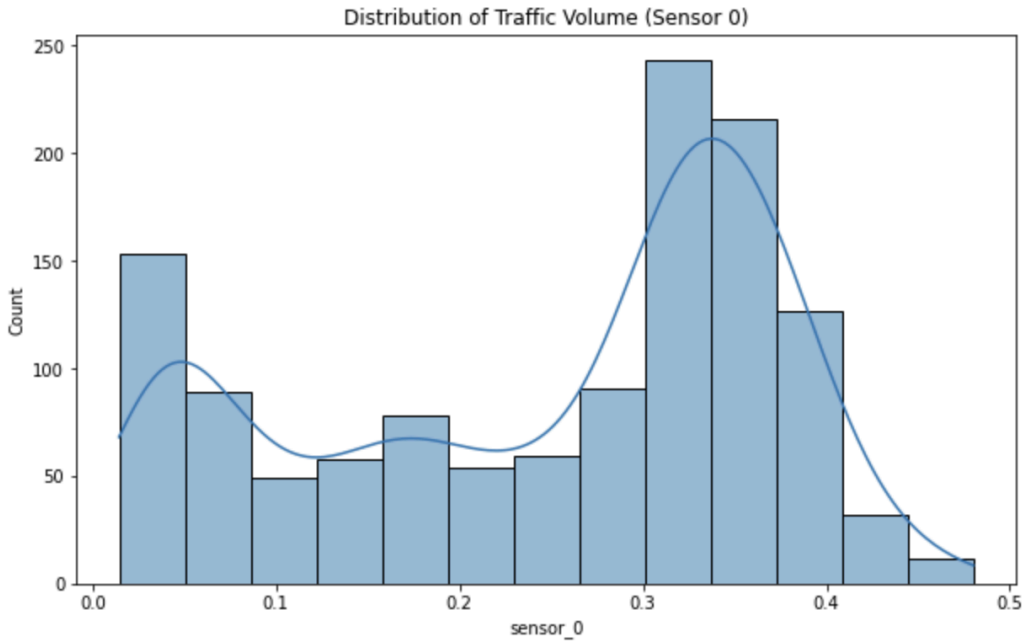
3.2 Autocorrelation Analysis

- **Autocorrelation (ACF) Plot**: The ACF plot shows significant positive autocorrelation at short lags, gradually decreasing as the lag increases. This indicates that traffic volume is strongly correlated with its recent past, suggesting regular periodicity. The slow decay of the ACF also points toward seasonality in the data.
- **Partial Autocorrelation (PACF) Plot**: The PACF plot shows a strong correlation at lag 1, which suggests that an autoregressive (AR) process may be a suitable model for this time series. Beyond lag 1, the PACF drops off quickly, indicating that immediate past values have the most influence on current traffic volumes.



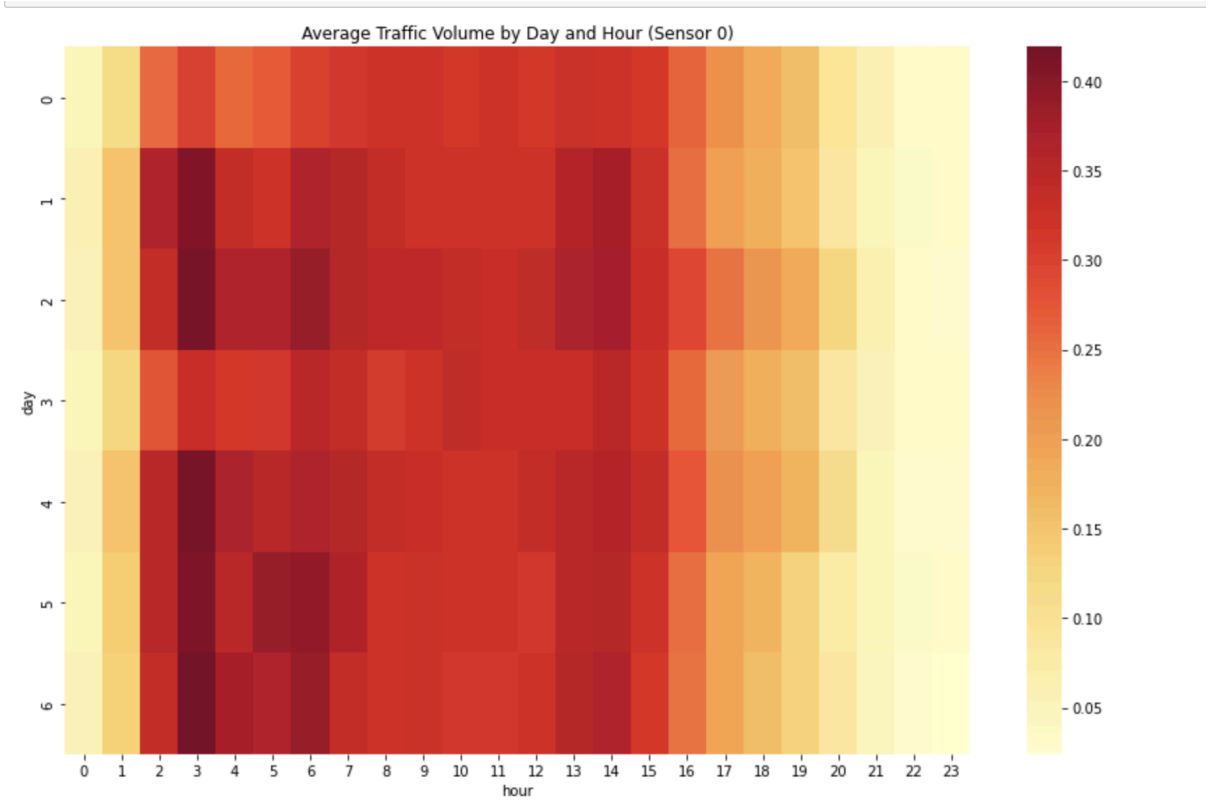
3.3 Traffic Volume Distribution

- **Distribution of Traffic Volume (Sensor 0)**: The histogram shows a right-skewed distribution, where most traffic volumes fall within the lower end of the range (between 0.0 and 0.3), but there are also instances of higher traffic volumes. The presence of a long tail to the right indicates occasional periods of significantly higher traffic flow, possibly during peak hours or special events.



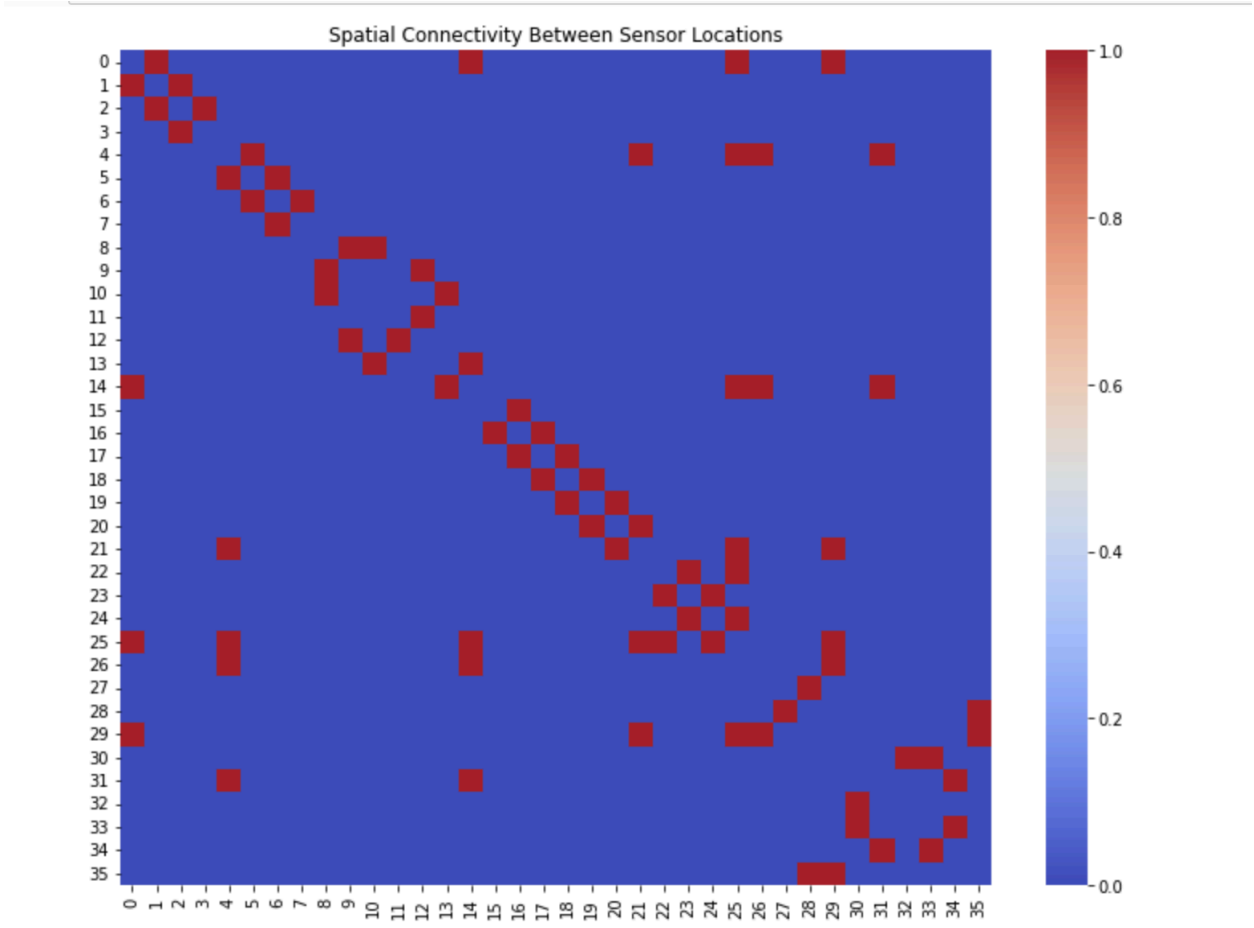
3.4 Temporal Patterns

- **Average Traffic Volume by Day and Hour (Sensor 0)**: The heatmap shows clear temporal patterns in traffic volume. Peak traffic tends to occur during specific hours of the day, with the highest volumes likely corresponding to rush hours (early morning and late afternoon). Days of the week also affect traffic, as evidenced by the variations in traffic flow across different days. Weekdays may exhibit higher traffic volumes, particularly during working hours, compared to weekends.



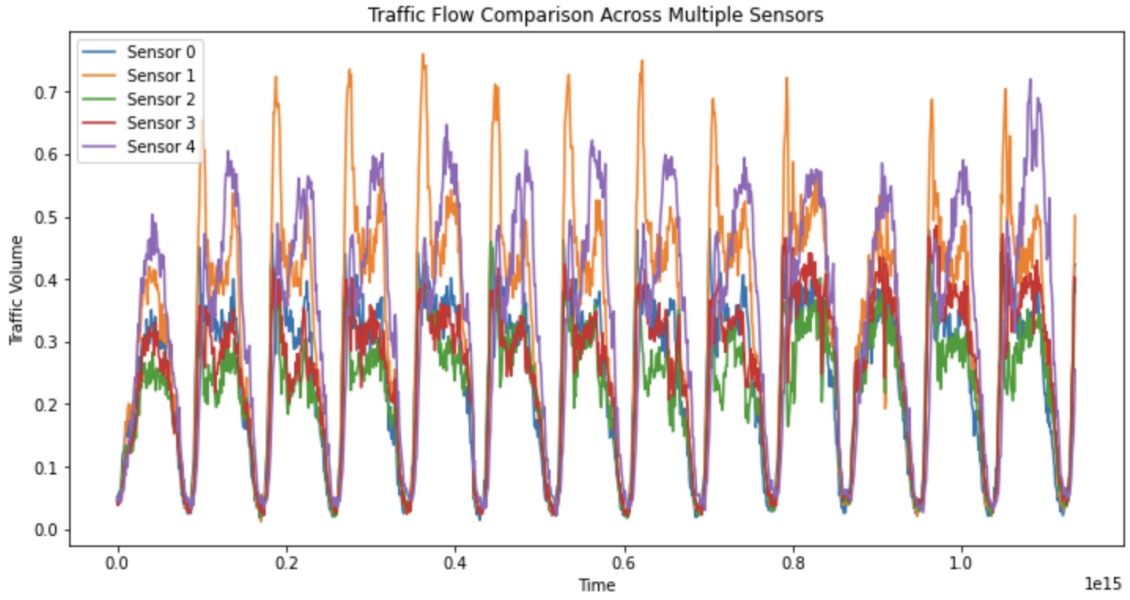
3.5 Spatial Analysis

- **Spatial Connectivity Between Sensor Locations**: The spatial connectivity heatmap shows the correlations between traffic sensors in different locations. Sensors with high positive correlation are clustered together, suggesting that traffic conditions in these areas are closely linked. This may indicate that these locations share traffic routes or are part of a larger connected road network. The absence of correlation (blue) between certain sensors indicates independent traffic flows between those locations.



3.6 Multi-Sensor Comparison

- **Traffic Flow Comparison Across Multiple Sensors**: The plot comparing traffic flows across multiple sensors shows that while the overall traffic flow pattern is similar across sensors (with all sensors exhibiting cyclical behavior), the magnitude of the traffic volumes varies between sensors. Some sensors detect consistently higher traffic volumes (e.g., Sensor 1 and Sensor 2), while others show lower traffic volumes (e.g., Sensor 0). This likely reflects differences in the locations of the sensors—some may be placed on busier roads or highways, while others may monitor less congested areas.



Key Features for Linear Regression:

1. **Time-Based Features**:
    - **Hour of the Day**: Captures daily patterns (e.g., rush hours).
    - **Day of the Week**: Accounts for weekday vs. weekend differences.
    - **Is Weekend**: Binary feature that differentiates weekends from weekdays.
    - **Month/Quarter**: Helps capture seasonal variations in traffic patterns.
2. **Rolling Averages**:
    - **3, 6, and 12-period rolling means**: Capture recent trends and smooth out short-term fluctuations, which provide stable input for the regression model.
3. **Lagged Traffic Flow (Optional)**: Adding lagged traffic values can help capture autocorrelation and improve predictive performance.

Insights for Model Selection:

- **Linear Regression**: The clear linear relationships between traffic flow and time-based features make linear regression a suitable baseline model. The use of time-based features and rolling averages should improve the model's performance by capturing cyclical and seasonal patterns.
- **Cyclical Nature**: Given the periodic behavior in traffic flow, more complex models (like **SARIMA** or **Exponential Smoothing**) could be explored later to capture non-linear relationships.