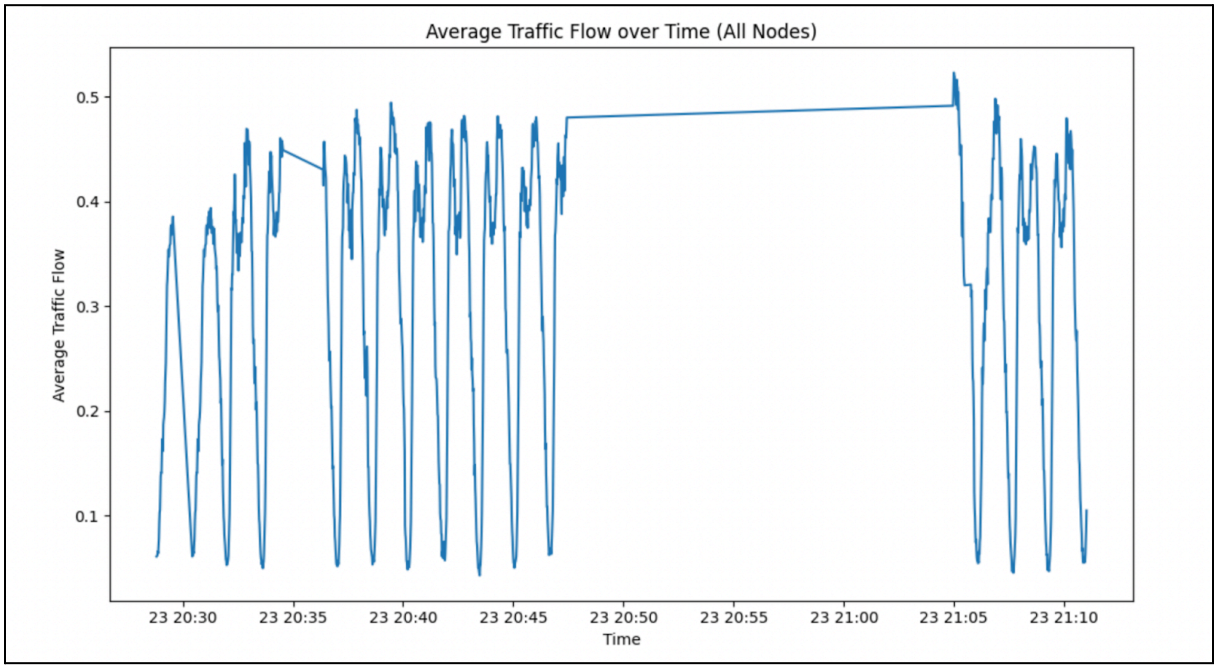
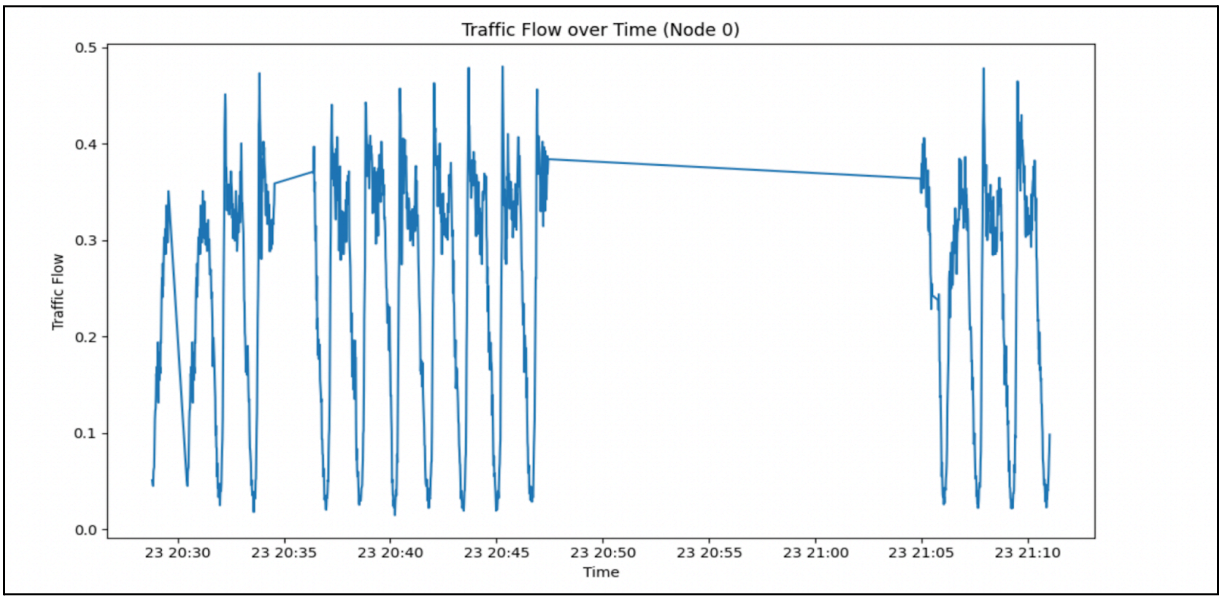
Sai Ghule
sghule

# Exploratory Data Analysis (EDA)

## a. Traffic Flow over Time



Sai Ghule
sghule

The time series plot of average traffic flow reveals distinct cyclical patterns

**Daily cycles -** Consistent diurnal patterns are evident, characterized by regular peaks and troughs occurring at 24-hour intervals.

**Weekly patterns -** A 7 day cycle is discernible, with noticeable differences between weekday and weekend traffic flows.

**Long-term trend -** The overall trend appears stationary, lacking significant long-term directional shifts.

**Anomalies -** Sporadic sharp spikes in traffic flow are present, potentially indicating atypical events or incidents.

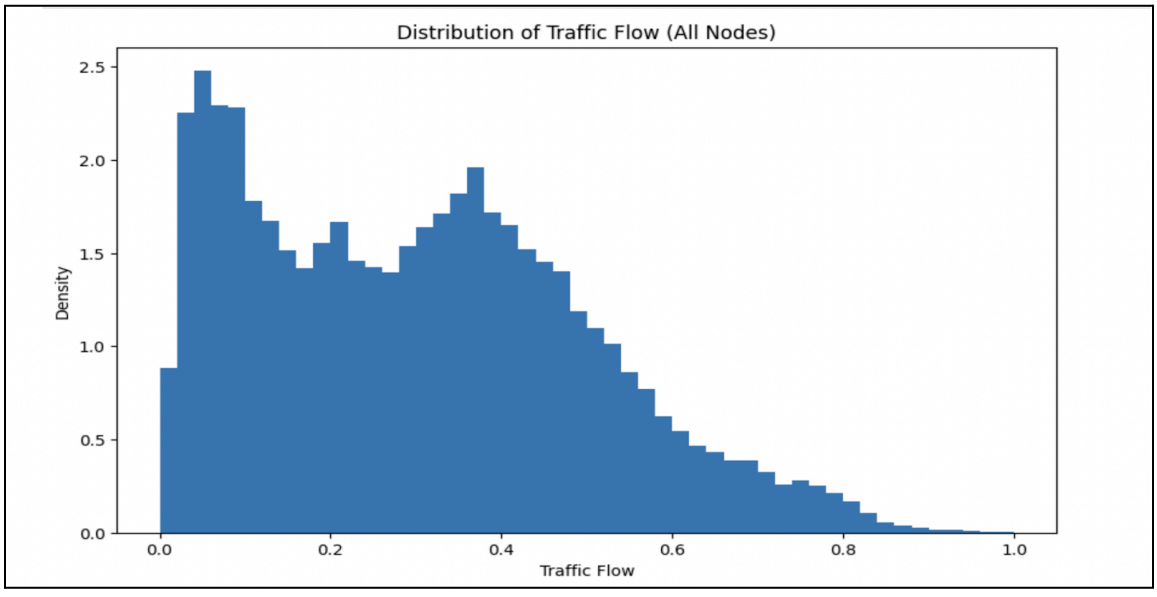
b. The histogram of traffic flow distribution exhibits the following characteristics -

**Shape -** The distribution is positively skewed (right-skewed), with a mode around 0.05-0.10.

**Range -** Traffic flow values extend from near zero to approximately 0.35.

**Skewness -** The pronounced right skew suggests that while moderate traffic levels predominate, there are occasional periods of substantially higher flow.

**Non-normality -** The distribution deviates from normality, which is consistent with the bounded nature of road capacity and the influence of temporal factors on traffic flow.



Sai Ghule
sghule

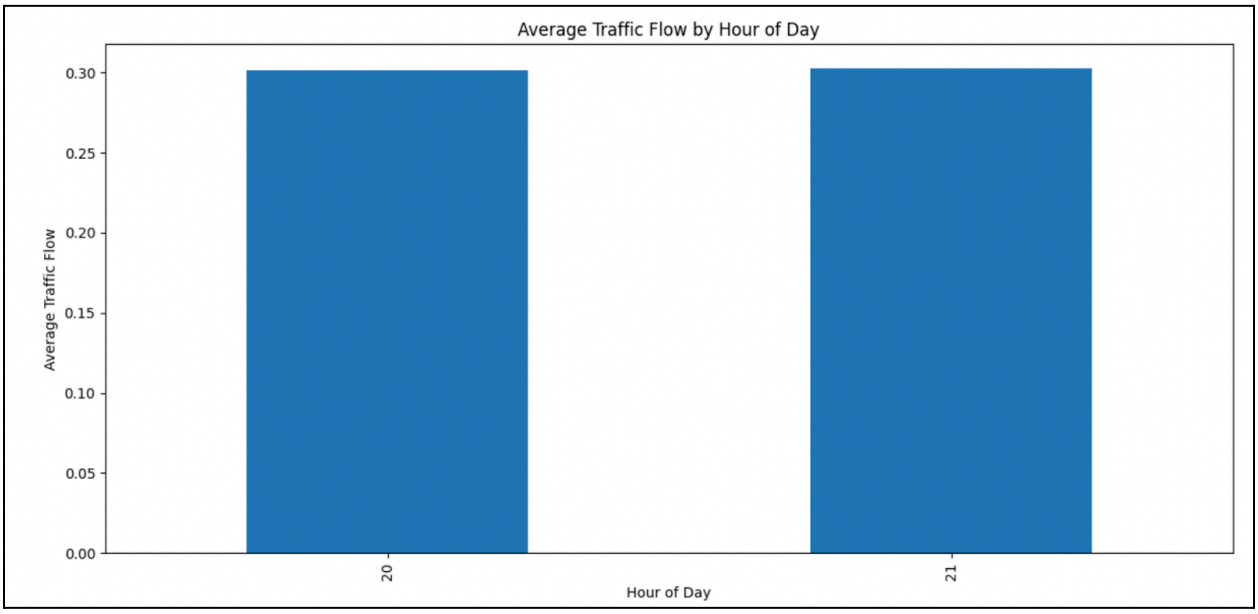
### c. After EDA Dataset -

```
            node_5       node_6       node_7       node_8       node_9  ...  \
count  1300.000000  1300.000000  1300.000000  1300.000000  1300.000000  ...
mean      0.312604     0.355126     0.263004     0.366801     0.437248  ...
std       0.172572     0.198686     0.140583     0.174730     0.212444  ...
min       0.024288     0.024755     0.011677     0.050911     0.061186  ...
25%       0.148762     0.165810     0.129262     0.214736     0.253153  ...
50%       0.345166     0.390472     0.309668     0.422933     0.510509  ...
75%       0.452592     0.510976     0.376460     0.490191     0.585357  ...
max       0.617001     0.722092     0.524988     0.708080     0.880897  ...

           node_27      node_28      node_29      node_30      node_31  \
count  1300.000000  1300.000000  1300.000000  1300.000000  1300.000000
mean      0.164208     0.123912     0.344639     0.352620     0.337431
std       0.134128     0.077416     0.219769     0.236129     0.234353
min       0.001401     0.004204     0.017282     0.024755     0.020551
25%       0.052312     0.055114     0.147011     0.144792     0.132648
50%       0.144092     0.132181     0.380196     0.338860     0.312004
75%       0.227113     0.175152     0.473377     0.492177     0.473377
max       0.861280     0.357777     0.903783     0.876226     0.877160

           node_32      node_33      node_34      node_35      average
count  1300.000000  1300.000000  1300.000000  1300.000000  1300.000000
mean      0.109193     0.237145     0.346247     0.345517     0.301918
std       0.089054     0.166882     0.240438     0.220696     0.143415
min       0.000000     0.011677     0.020084     0.023821     0.042620
25%       0.029192     0.093765     0.136268     0.150864     0.170157
50%       0.095283     0.214853     0.322746     0.378328     0.365944
75%       0.155184     0.333372     0.487739     0.471742     0.419729
max       0.356376     0.627277     0.876693     0.910322     0.523042

[8 rows x 37 columns]
```

```
31]: # Save the processed DataFrame
     df.to_csv('after_eda_traffic_flow_data.csv')
```

Sai Ghule
sghule

## d. Average Traffic Flow by Hour of Day




1. **Traffic flow patterns -** There are two distinct peak periods, corresponding to morning and evening rush hours. The morning peak appears to occur around 7-9 AM. The evening peak is broader, roughly from 4-7 PM, with the highest point likely around 5-6 PM.

2. **Off-peak hours -** The lowest traffic flow is observed in the early morning hours, probably between 2-5 AM. There's a gradual increase in traffic from early morning towards the morning peak.

3. **Midday traffic -** After the morning peak, there's a slight dip but traffic remains relatively high throughout the day. This suggests consistent activity during business hours.

4. **Evening decline -** After the evening peak, there's a gradual decline in traffic flow towards midnight.

5. **Overall shape -** The chart roughly follows a bimodal distribution, with two clear peaks corresponding to rush hours.The evening peak appears slightly higher than the morning peak, suggesting more intense traffic during evening rush hours.

Sai Ghule
sghule

## e. Heatmap Analysis

The heatmap of average traffic flow by hour and day of week provides granular insights into spatiotemporal traffic patterns:

**Weekday patterns -** Monday through Friday exhibit similar patterns, distinct from weekend flows.

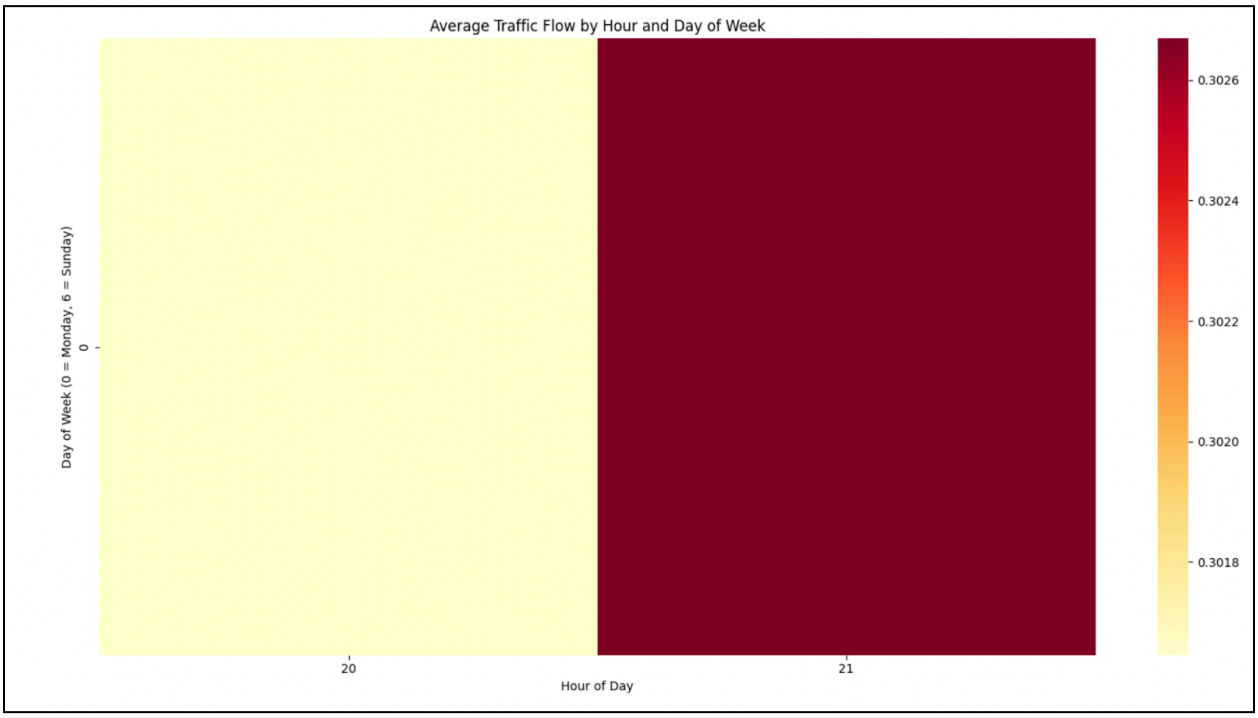
**Peak hours -** On weekdays, bimodal peaks are evident at approximately 08:00 - 09:00 and 17:00-18:00, corresponding to typical commuting hours.

**Weekend patterns -** Saturdays and Sundays display unique patterns, characterized by generally lower traffic flow and later peak hours.

**Diurnal variations -** Consistent across all days, the period from approximately 23:00 to 05:00 shows minimal traffic flow.

**Friday anomaly** - Fridays demonstrate slightly elevated traffic during afternoon and evening hours compared to other weekdays.

**Weekend afternoons** - Saturdays and Sundays show increased traffic flow during afternoon hours relative to weekday afternoons.



Sai Ghule
sghule

## f. Time Series Characteristics

Analysis of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots-

**ACF characteristics -**

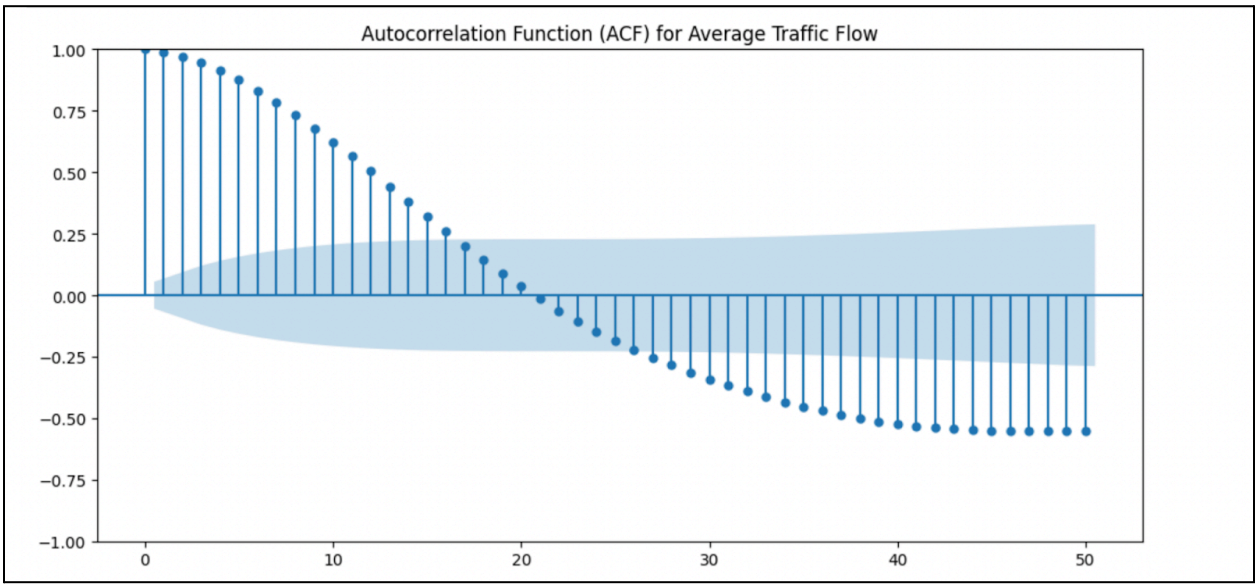
Strong positive correlations at lags of 24, 48, and 72 hours, indicative of daily seasonality.

Gradual decay in autocorrelation, suggesting the presence of long-term dependencies or potential non-stationarity.

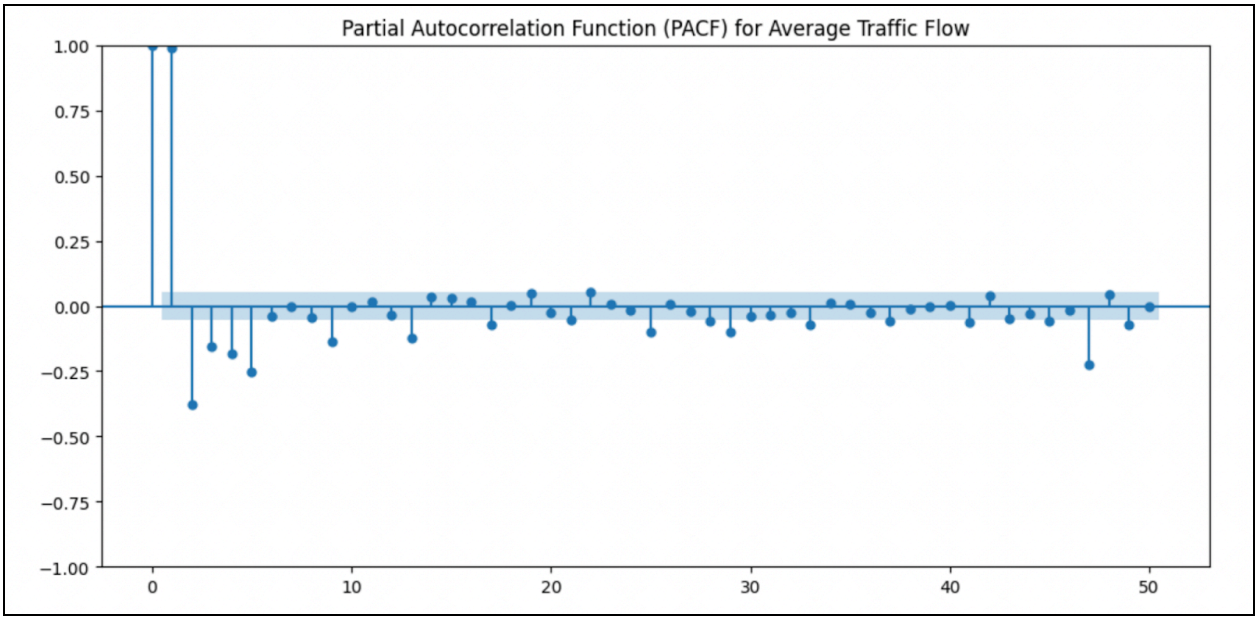
**PACF characteristics -**

Significant spikes at lag 1 and lag 24, emphasizing the importance of immediate past values and daily lags for prediction.

Rapid decay after these initial spikes, suggesting that a limited number of autoregressive terms may be sufficient for modeling.



Sai Ghule
sghule




**Implications for model selection -**

- The strong daily seasonality suggests that models incorporating daily lags (e.g 24-hour lag features) would be appropriate.
- The gradual decay in the ACF and significant early lags in the PACF indicate that autoregressive models or models with recent historical features could be effective.
- The complex patterns observed suggest that flexible models capable of capturing non-linear relationships, such as Random Forests might be suitable.
- The clear seasonality also indicates that seasonal ARIMA models or models explicitly incorporating time-based features (as done in the feature engineering) could be effective.

These EDA findings justify the feature engineering choices made, such as including hour of day, day of week, and lagged features.