# Capstone Three Final Report - Hourly Traffic Volume Forecasting

## 1. Problem Statement

Accurate short-term forecasting of hourly traffic volume is critical for operational traffic management, infrastructure utilization, and congestion mitigation. Transportation agencies require reliable forecasts to allocate resources, manage peak demand, and respond proactively to expected traffic conditions. This project aims to develop and evaluate time-series forecasting models capable of predicting hourly traffic volume over short horizons, with a focus on interpretability and operational reliability.

## 2. Dataset Description

The dataset used in this analysis is the Metro Interstate Traffic Volume dataset, which contains hourly traffic counts collected from a major metropolitan interstate corridor. The target variable is traffic_volume, representing the total number of vehicles observed per hour. Explanatory variables include timestamp information, weather conditions (temperature, precipitation, cloud coverage), and holiday indicators. Initial inspection revealed structural issues such as duplicate timestamps, missing hourly intervals, and implicit encoding of non-holiday periods.

## 3. Data Wrangling

Before any analysis or modelling could be performed, the raw dataset required structural correction to ensure temporal consistency and suitability for time-series modelling.

3.1 Aggregating Duplicate Timestamps

The original dataset contained multiple observations for the same hourly timestamp, which violates the assumption of one observation per time step required for time-series analysis.

To resolve this, the data was aggregated at the hourly level, ensuring exactly one row per date time:

- Numeric features (e.g., temperature, precipitation, traffic volume) were aggregated using the mean, representing average conditions within the hour.
- Categorical features (e.g., weather conditions, holiday indicators) retained the most frequent (mode) value, representing the dominant condition during that hour.

This aggregation step removed duplicate timestamps while preserving the underlying signal in both continuous and categorical variables.

This step ensures structural validity of the time series and prevents bias caused by uneven intra-hour sampling.

### 3.2 Enforcing a Continuous Hourly Time Index

After aggregation, the time series was reindexed to a continuous hourly frequency. This explicitly introduced missing timestamps where no observations were originally recorded, making temporal gaps visible rather than implicitly ignored.

This step is critical for:

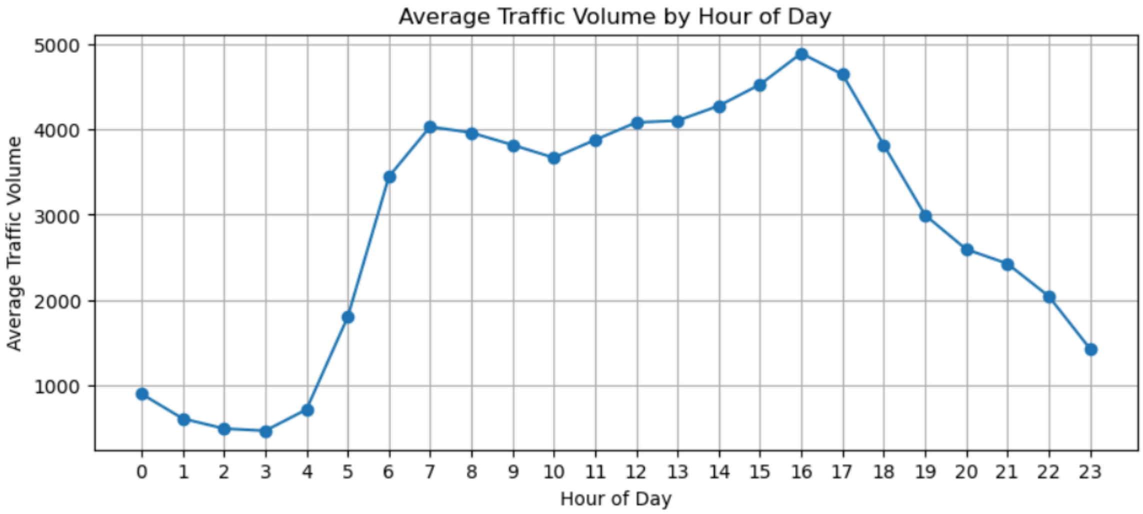
- detecting missing periods
- enabling lag-based features

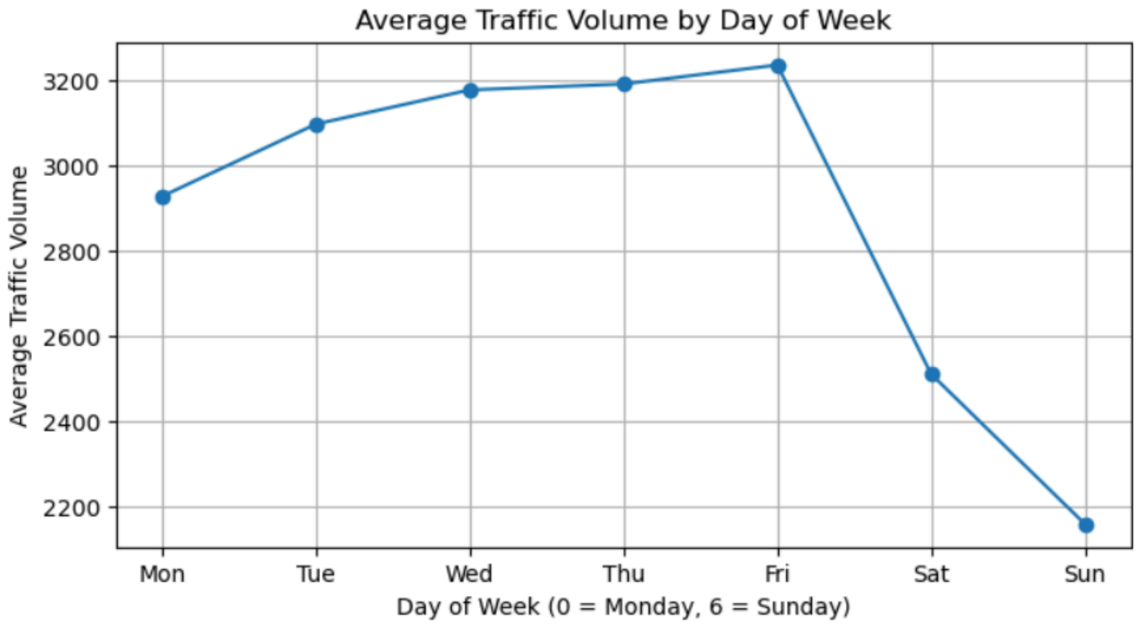
### 3.3 Standardizing Holiday Encoding

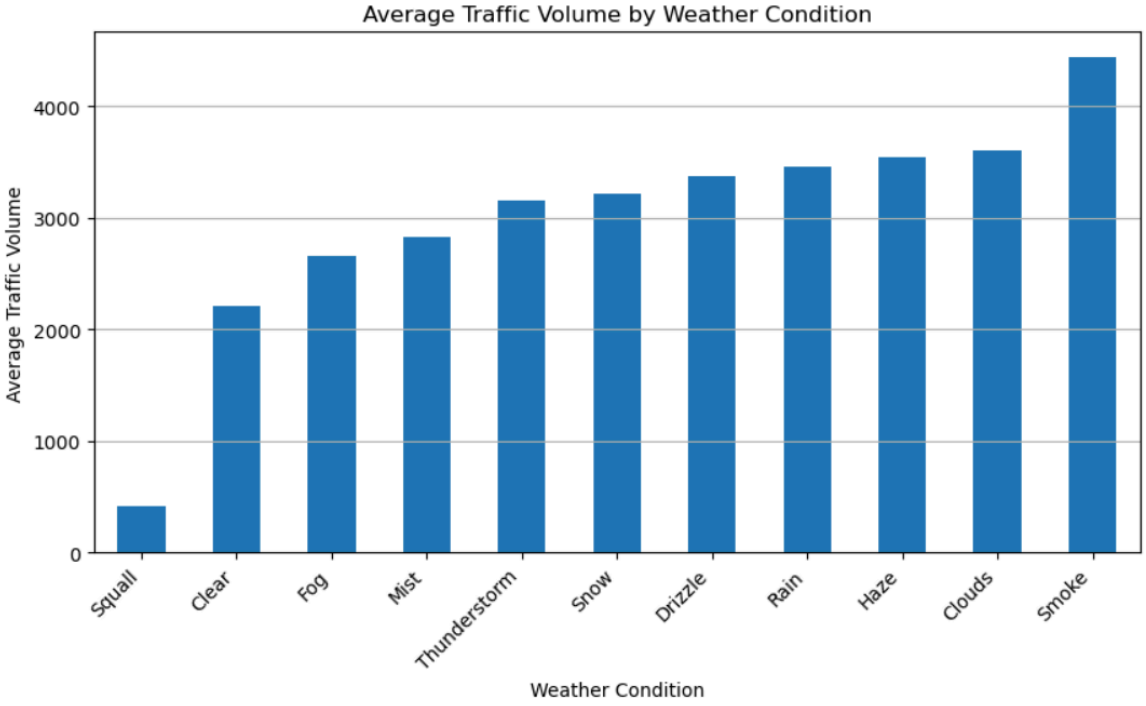
Holiday values originally encoded as missing were reinterpreted as non-holiday periods and standardized accordingly.
This ensured consistent semantic meaning across the dataset and prevented misinterpretation of missing values as unknown states.

## 4. Exploratory Data Analysis (EDA)

Exploratory analysis revealed strong and highly regular daily and weekly seasonality in traffic volume. Morning and evening commute peaks dominate weekday traffic patterns, while weekends exhibit consistently lower volumes. Weather conditions such as heavy rain and snow are associated with reduced traffic levels, though these effects are secondary to temporal patterns. Overall, the traffic-generating process appears stable over time, supporting the use of time-series forecasting models.







## 5. Data Preprocessing

Following exploratory analysis, the dataset was prepared for modeling. Short gaps introduced during reindexing were forward-filled to preserve realistic temporal continuity. All modeling variables were converted to appropriate numeric formats, and the dataset was split into training and test sets using a time-based split to avoid data leakage. This step ensured that models were trained exclusively on historical data and evaluated on unseen future observations.

## 6. Feature Engineering

Feature engineering focused on constructing interpretable and computationally efficient inputs for time-series models. Holiday indicators were encoded into binary variables, and weather variables were selected based on domain relevance. For SARIMAX models, the feature set was intentionally simplified to avoid high-dimensional dummy variables while preserving meaningful external signals.

## 7. Modeling Strategy

A progressive modeling strategy was adopted, starting with a simple seasonal naive baseline and advancing to more complex statistical models. This approach aligns with forecasting best practices by establishing a strong benchmark before introducing additional complexity. The strategy also satisfies the requirement to evaluate multiple models while maintaining interpretability.

## 8. Model Implementation

Five time-series models were implemented to evaluate progressively more complex forecasting approaches while maintaining fair, time-aware comparisons. All models were trained and evaluated using a consistent train–test split based on time order to avoid data leakage.

The models were implemented in the following sequence:

1. Seasonal Naive Baseline
   A rule-based benchmark that predicts traffic volume using the observed value from the same hour one week earlier.

2. SARIMA (No Seasonality)
   A non-seasonal autoregressive model capturing only short-term temporal dependence.

3. SARIMA (Daily Seasonality)
   A seasonal model with explicit 24-hour seasonality to capture intraday traffic patterns.

4. SARIMAX (Full Exogenous Features, No Seasonal State)
   A model incorporating weather and holiday variables as exogenous regressors, without explicit seasonal structure.

5. SARIMAX (Daily Seasonality + Light Exogenous Features)
   A seasonal SARIMAX model combining daily seasonality with a reduced set of continuous weather variables and a binary holiday indicator.

## 9. Model Comparison and Metrics

Model performance was evaluated using Root Mean Squared Error (RMSE) on a held-out test set. Lower RMSE values indicate better out-of-sample predictive accuracy.

Model Performance Results:

1. Model: Seasonal Naive Baseline
   RMSE: 645.64

2. Model: SARIMA (No Seasonality)
   RMSE: 3857.64

3. Model: SARIMA (Daily Seasonality)
   RMSE: 982.74

4. Model: SARIMAX (Full Exogenous, No Seasonality)
   RMSE: 2015.26

5. Model: SARIMAX (Daily Seasonality + Light Exogenous)
   RMSE: 957.92

## 10. Final Model Selection

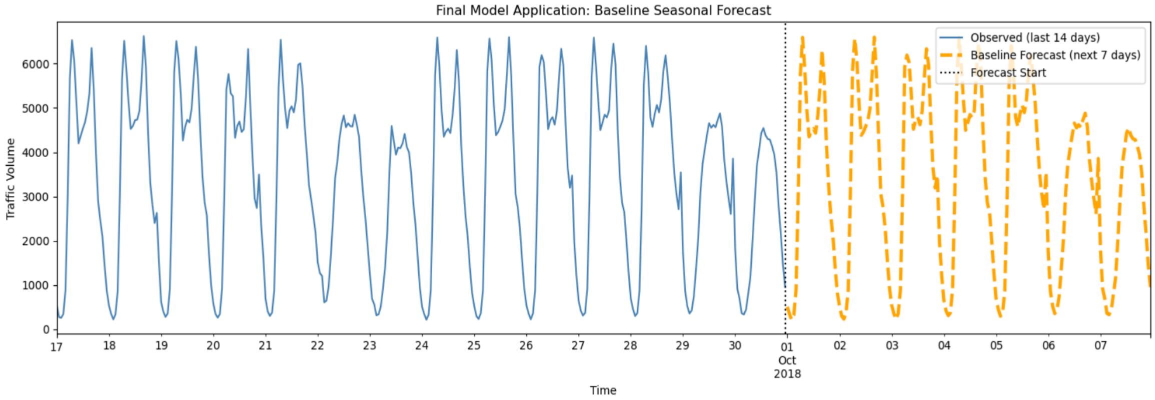
Based on the RMSE comparison, the Seasonal Naive Baseline was selected as the final model for short-term forecasting. This model achieved the lowest RMSE (645.64) on the held-out test set, substantially outperforming all SARIMA and SARIMAX variants.

Although more complex models incorporating autoregressive structure and exogenous variables were evaluated, none provided a meaningful improvement in predictive accuracy. In several cases, additional complexity led to significantly higher error, indicating overfitting and unnecessary parameter estimation.

Given its superior performance, interpretability, and minimal computational cost, the Seasonal Naive Baseline provides the most reliable and practical solution for short-horizon hourly traffic forecasting in this dataset.

## 11. Final Model Application

The selected baseline model was applied to generate a real, future-oriented forecast for the next seven days (168 hours) beyond the observed dataset. The forecast demonstrates continuity with recent history, preserves daily and weekly patterns, and remains within realistic historical bounds. This confirms the model's suitability for short-term operational forecasting.



## 12. Business Recommendations

The analysis indicates that short-term hourly traffic volume is highly predictable and dominated by stable daily and weekly seasonal patterns. Based on model comparison results, the seasonal naive baseline model—using traffic from the same hour one week earlier—should be adopted as the default forecasting approach for horizons up to seven days. This model achieved the lowest prediction error (RMSE ≈ 646), outperforming more complex SARIMA and SARIMAX models, which introduced additional estimation noise without improving accuracy.

Second, traffic management teams can use these forecasts to proactively plan staffing, monitoring, and congestion-mitigation activities around predictable peak periods. Both historical data and the final forecast show that morning and evening commute peaks occur consistently at the same hours, with higher volumes on weekdays than weekends. This regularity allows operational resources and maintenance activities to be scheduled with confidence during expected high- and low-demand periods, improving efficiency while minimizing congestion risk.

Third, weather conditions and holidays should be treated as monitoring and exception signals rather than core forecast drivers. Although exploratory analysis showed that adverse weather and holidays can affect traffic levels, incorporating these variables into statistical models provided only marginal improvements and did not outperform the seasonal baseline. From a business perspective, weather and holiday indicators are best used to trigger manual review or alerts during extreme conditions, rather than increasing model complexity for routine forecasting.

## 13. Limitations and Future Work

While effective for short-term forecasting, the baseline model cannot account for unexpected events such as accidents or large public gatherings. Future work could incorporate real-time incident data, event calendars, or adaptive models that update dynamically as new data arrives. Exploring probabilistic forecasts could also provide uncertainty estimates to support risk-aware decision-making.