Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Analyzing traffic patterns of M50 and Traffic Prediction

Ritwik Kumar

April 17, 2023

SUPERVISOR: Prof. Siobhan Clarke

A dissertation submitted in partial fulfillment of the requirements for the degree of
BAI (Computer Engineering)

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this or any other University, and that unless otherwise stated, is my own work.

Signed: Ritwik Kumar Date:17/04/2023

Abstract

This paper describes an effort to study and analyze the traffic flow patterns on M50 a major motorway in Ireland, and predict the traffic congestion pattern through a Long Short-Term Memory (LSTM). Recurring patterns in traffic flow were deduced from the analysis of traffic data through various statistical tools and machine learning algorithms. The study shows that there are significant variations in traffic flow patterns across different time periods and days of the week. The traffic congestion model was predicted by training the LSTM model on the historical traffic data NorthBound 2 lane in order to predict congestion on the Highway. The model demonstrated very low loss during the epochs which means it has high accuracy in predicting congestion levels which can be used to alert commuters and transport authorities. This research is a way forward in the development of effective strategies and solutions to manage and control traffic congestion on major motorways, which can improve commuter safety and cut down on travel time.

Acknowledgments

I would like to express my sincere gratitude to Siobhan Clarke, my project supervisor, for her unwavering support and guidance throughout the course of this project. Without her assistance, this project would not have been possible.

I would also like to extend my appreciation to Aqeel Kazmi, who provided invaluable feedback during the early stages of setting up my project. Your input and suggestions were incredibly helpful and greatly contributed to the success of this project.

Once again, thank you both for your support and contributions to this project.

Contents

List of Figures

Nomenclature

EDA- Exploratory Data Analysis
RNN-Recurrent Neural network
CNN-convolution Neural network
HGVs-Heavy Goods vehicle(vehicles in this thesis is HGVs only)
LSTM- Long-short Term Memory
TII- Transport Infrastructure Ireland
ARIMA - Autoregressive Integrated Moving Average

1 Introduction

Traffic congestion is a major issue in urban areas. It occurs when there are too many cars present on the road and as a result, the flow of vehicles gets slowed down or comes to rest. The number of vehicles on the road has increased significantly along with the demand for transportation in recent years, putting pressure on the infrastructure for highway traffic and contributing to a number of congestion issues. The M50 is one of Ireland's busiest highways, and during rush hour it is severely congested with traffic. This thesis's objective is to examine the M50's traffic congestion patterns and create a model for predicting congestion using the Long Short-Term Memory (LSTM) algorithm.

To solve this problem, it is necessary to have an exploratory data analysis(EDA) of the historical data provided by TII, which will help in finding out various trends, and patterns of traffic. With the help of the analysis and applying further traffic theory we'll be able to build a Traffic flow prediction model. Time series data, like traffic flow, can be effectively modeled using LSTM which is a type of recurrent neural network. We can more accurately predict congestion by utilizing LSTM to capture the intricate patterns of traffic flow.

The findings of this study will shed important light on various patterns of traffic congestion on the M50 due to HGVs and how well the LSTM algorithm predicts congestion. In order to reduce traffic congestion on the M50 and enhance Dublin's overall transportation system, transportation planners and policymakers will find the findings to be helpful.

1.1 Motivation

Dublin is considered the second-worst city in terms of traffic in Europe. According to statistics, the typical Dublin driver spends 250 hours a year stuck in traffic. On the highways and in metropolitan areas, it is a serious issue that needs to be addressed in order to make a better transportation system in Dublin.

The 45-kilometer-long M50 highway encircles Dublin and links all important thoroughfares to the city's core. Every day thousands of commuters and business vehicles use it as an essential component of Ireland's transportation system. However, the high volume of traffic during peak hours frequently causes congestion, which lengthens travel times, increases fuel use, and pollutes the air. In this project our main focus was on heavy goods vehicles (HGVs)

related to traffic congestion on the M50 motorway which not only affects the transportation sector but also the environment and the whole economy.

Many cities around the globe struggle with serious traffic problems and many people suffer because of congestion which results in the aggressive behavior of drivers which may result in aggressive actions. The motivation of this thesis is to analyze the traffic patterns from EDA and develop a traffic congestion prediction model that will accurately forecast traffic congestion and help traffic planners and authorities to take steps in advance in order to tackle the situation better. This can be achieved using the machine learning algorithm and the application of LSTM neural networks. The main aim of this research is to provide an effective solution for traffic congestion caused by HGVs on the M50 highway which will eventually solve the problem of emissions.

1.2 Problem Statement

Heavy goods vehicles (HGVs) produce a large amount of traffic congestion on the M50 highway during peak hours, which prolongs travel times, increases fuel consumption, and pollutes the environment. Additionally, the congestion makes it impossible for emergency vehicles to pass and increases traffic infractions, red signal running, breakdowns, and accidents that may result in personal injury or property damage. The businesses operating in the nearby area experience a loss in productivity and an increase in transportation expenses as a result of the M50's congestion.

As we are primarily concentrating on the effects of the congestion caused by HGVs on the M50 freeway in this thesis, the following is a discussion of some of the issues brought on by HGVs.

- Congestion – Heavy Good Vehicles (HGV) is one of the big reasons for traffic congestion on highways as HGV vehicles are very large in size, take more space on the road, and even travel at slow speed as compared to normal vehicles which leads to more fuel consumption.

- Emissions-HGVs are a substantial source of emissions, which include carbon dioxide and nitrogen oxides, two of the main pollutants of the air. Because diesel engines, which make up the majority of HGVs, are less efficient, they consume more fuel, which increases carbon dioxide emissions and other environmental pollutants.

The issue is made worse by the absence of reliable techniques for precisely studying the causes of HGV congestion on the M50 and forecasting congestion. Due to the intricate patterns of HGV traffic flow, current methods for predicting congestion have difficulty effectively predicting HGV congestion.

1.3 Objective

The ultimate goal of this project is to Analyze(EDA) the traffic pattern of HGVs on M50 highways and determine various insights from the provided data. The further aim is to develop an LSTM-based traffic flow prediction model that will predict the traffic flow on the

basis of historical data that has been provided by Transport Infrastructure Ireland(TII). This will help transportation planners and operators to make informed decisions regarding traffic management and reduce the negative impacts of congestion such as increased travel times, higher fuel consumption, and carbon emission.

1.4 Thesis Layout

Overview of thesis and a short description of all the chapters:

- Chapter 2 gives information about the background of the project and the related work and research done for this project.
- Chapter 3 gives information about the design and implementation of the data analysis and model creation.
- Chapter 4 gives information about various results and insight obtained from Analysis and model creation.
- Chapter 5 will provide a conclusion and talk about the difficulties faced and future scope

2 Background and Related Work

Traffic congestion is a significant problem in urban areas, and traffic pattern analysis and prediction have become crucial research fields in recent years. Initially the majority of emphasis was concentrated on creating methods that could be used to simulate traffic characteristics including volume, density, and speed, and then generate predicted traffic situations, which might be seen as classical techniques[1]. Some representative studies show that a variety of machine learning algorithms can be used for forecasting traffic flow which will provide more efficient results as compared to the traditional methods.

In this section, we are discussing the background of the project and explain important terms related to the project. Further discussing various analyzing methods that have been used by the researchers, various traffic flow prediction models and their architectures used, and the shortcomings of their research and results.

2.1 M50 Highway

fig-2.1 Map of M50 Highway

The M50 highway is one of the most important highways in Dublin, Ireland connecting many important routes around the city and acting as a crucial link in the country's motorway system.It has a total length of approximately 45 kilometers and is one of the busiest roads in the nation. It sees a heavy traffic load of about 145,000 vehicles each day. Many researchers are working on M50 as the problem of congestion is increasing day by day. So my aim is to study the data of M50 and gather useful insights from its traffic patterns.

2.2 Heavy goods vehicles(HGVs)

Our main focus is on traffic created by HGVs. HGVs are large commercial vehicles used for the transportation of goods. HGVs are primarily employed for long-distance. They are of two types - HGV Rigid and HGV Articulated.

fig-2.2 Heavy good vehicle

HGV Rigid - A tractor-trailer, sometimes referred to as a heavy goods vehicle and is composed of a cabin unit and a separate trailer unit. The cab unit includes the engine, the driver's cabin, and the controls; the cargo is held in the trailer unit, which is connected to the cab. Because of their vast size, they have plenty of room to carry a lot of stuff.

HGV Articulated – A heavy goods vehicle that consists of a cab unit and a semi-trailer unit is referred to as an HGV art, articulated lorry, or semi-trailer truck. The semi-trailer unit is connected to the cab and contains the cargo, while the cab unit has the engine, driver's compartment, and controls. It is almost similar in size compared to an HGV rig and hence can carry lots of stuff as well.

2.3 Transport Infrastructure Ireland

TII is a state agency whose main work is to manage the country’s national roads, road networks, railways, ports, and airports. TII is also responsible for the development of the transport system in the country. They even develop new ideas that will help in the betterment of Ireland's transport infrastructure.

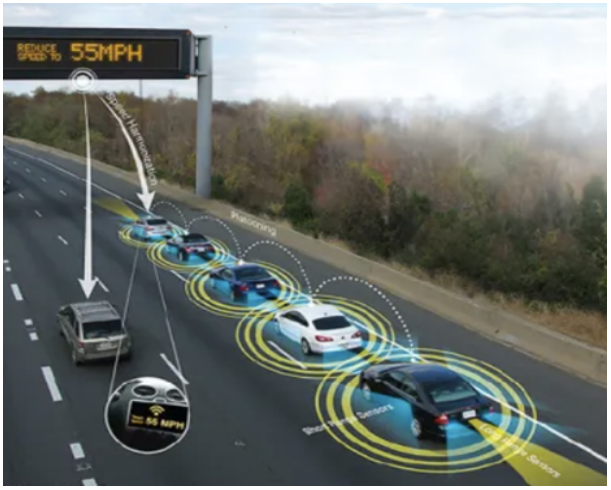

Fig-2.3 Traffic sensors

The data information is gathered by a series of sensors and cameras that are present on national roads and they continuously record the traffic data[5]. Traffic counters provide information based on the volume of traffic by hour of day and vehicle class. We can download the traffic dataset from the website of TII which is in the raw form based on the counter data.

2.4 Traffic Flow Terminology

Speed, flow, and density are significant factors that are utilized to analyze traffic situations in traffic theory. Through the basic traffic flow equation, they are connected to one another[3].

Traffic Speed: The rate at which cars move along a road segment is known as the traffic speed, and it is expressed in terms of the distance traveled in a certain amount of time (for example, in kilometers per hour). Due to the fact that it has an impact on traffic flow and density, it is a crucial element in traffic analysis[3].

Traffic flow- It is measured in units of vehicles per unit time (for example, vehicles per hour), and is the volume of cars that travel through a road section in a certain amount of time[3].

Traffic density (measured in terms of vehicles per unit length or vehicles per kilometer) is the number of vehicles using a certain length of a road at any given moment. k can be determined across a road section given ΔX length over a measurement interval at a specific time, like S1 as follows[3]:

$$k = \frac{n}{\Delta X}$$

n is the number of vehicles at a given time t on location ΔX.

The relationship between Speed, flow, and density is given below

$$q = k * v$$

where v is the traffic speed

q is the traffic flow (measured in vehicles per unit time)

k is the traffic density (measured in vehicles per unit length).

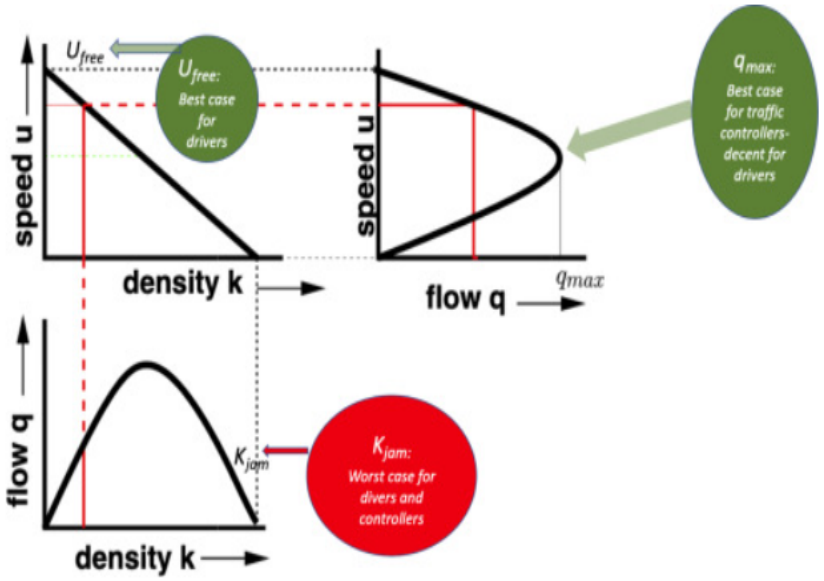

fig-2.4 Relationship between speed density and flow[2]

These variables are interdependent and affect each other in complex ways. For example, as traffic density increases, traffic speed typically decreases due to congestion. This in turn reduces traffic flow, which can further increase congestion. Understanding the relationships between speed, flow, and density is important in predicting and managing traffic conditions.

Relative traffic speed is also commonly used to quantify congestion. Relative speed is calculated as the ratio of a Traffic speed to the maximum Traffic speed on a particular road segment.

Relative speed = Traffic speed/max(Traffic speed)

Higher the relative traffic speed, the less the congestion.

The lower the relative traffic speed, the more the congestion.

2.5 Types of Congestion

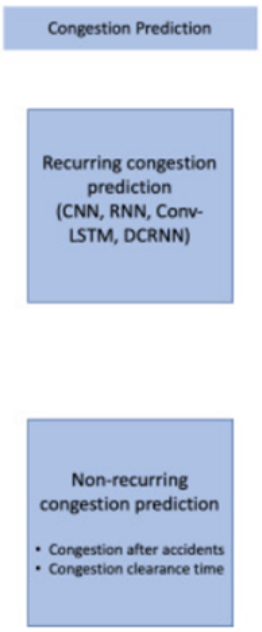

Fig-2.5- Types of congestion

There are two types of traffic congestion, recurring congestion and the other is non recurring congestion. Recurring congestion shows similar trends in traffic. Recurring congestion occurs when there is constant congestion in a particular region or at a particular time, such as during rush hour in a particular portion of the city or on a particular day of the week[1].

Whereas, non-recurring congestion does not show any similar pattern or trend. Non-recurring congestion occurs when there is an accident, road construction, natural disaster, or special events such as concerts that draw huge numbers of vehicles onto roads which might result in this kind of congestion[1]. Non-recurring congestion is more challenging to handle than recurring congestion since it might happen abruptly and is less predictable.

2.6 Deep learning

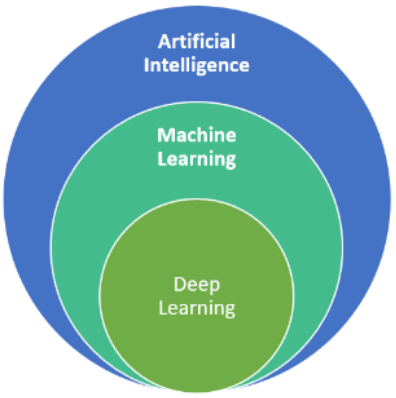

fig 2.6- Relationship between AI,ML,DL

Deep learning is a subset of machine learning that uses artificial neural networks to model and solve complex problems. It involves training a neural network with large amounts of data to learn and make predictions or decisions based on patterns in the data. it is used for a wide range of applications where traditional machine-learning algorithms approaches may not be sufficient to handle the complexity of the data. It has the ability to learn and make predictions based on patterns in large datasets making it a powerful tool for solving many real-world problems.

2.6.1 Neural Networks

A neural network is a machine learning algorithm that is based on the structure and operation of the human brain. Every neuron is a mathematical function that functions similarly to a biological neuron. It is a group of interconnected neurons or nodes that may learn to carry out challenging tasks by varying the strength of their connections[12]. Each neuron takes in information from other neurons, processes it, and then sends a signal to other neurons.

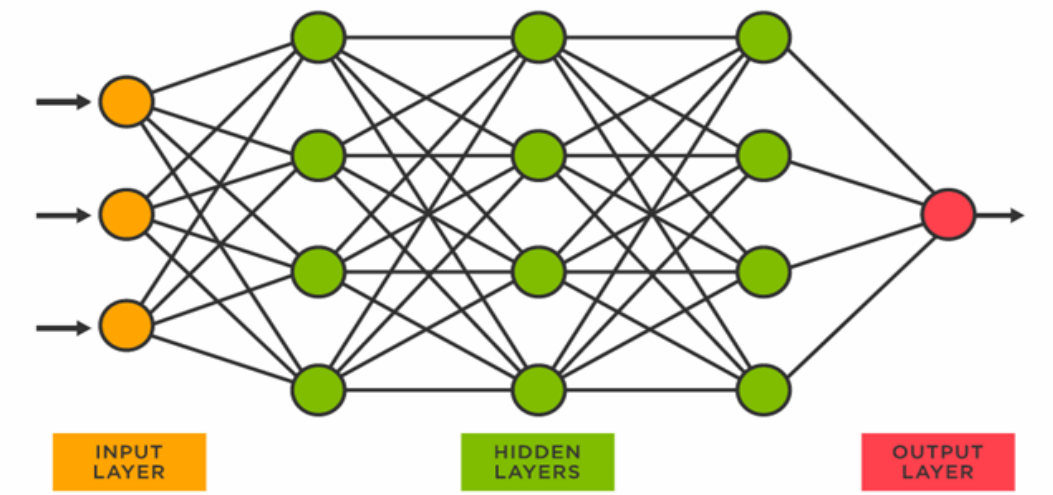

fig-2.7 Neural network

The input layer is the initial layer of the network, and the output layer is the final layer. The intermediate layers, referred to as hidden layers, carry out calculations to convert the input data into an output prediction[2]. During training, the network gets input data, and then the network modifies the weight of the connections between neurons to reduce the discrepancy between its anticipated output and the desired output. This process is performed many times until the network is able to accurately predict the goal output for new input data.

Although neural networks have numerous benefits, they also have drawbacks. The main drawback is we need to have large amounts of data for the training of neural networks.

2.6.2 Recurrent neural network

Recurrent neural networks (RNNs) are a special kind of neural network that takes into account both the current input and what it has learned from prior inputs as a result of feedback connections, making them especially helpful for tasks involving sequential input like time series prediction or natural language processing[1]. RNNs are well suited for processing sequential data, such as time-series data, because these feedback connections

enable RNNs to maintain an internal state from one time step to the next.

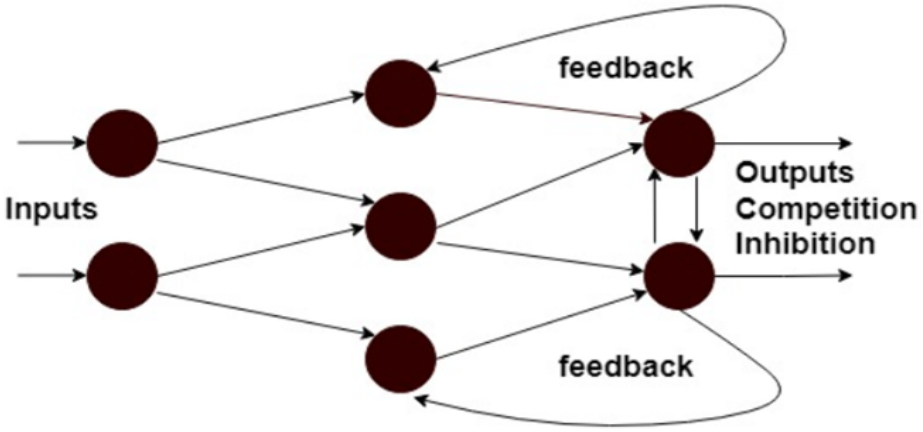

fig 2.8 - Recurrent neural network

The two major problems that RNNs face are the vanishing gradient problem and the exploding gradients problem. The exploding gradients issue arises when the gradients are too large which results in an unstable model. Whereas the vanishing gradient problem occurs when the network is trained on a significant amount of data because the gradient used to change how much each neuron is weighted grows too small, making it impossible for the network to learn and update its weights[2].

Recurrent Neural Networks (RNNs) are crucial in traffic flow prediction models as they are built to handle sequential data with temporal dependencies, a typical feature of traffic flow data. RNNs are excellent at capturing long-term relationships and time-varying patterns in traffic data, which makes them suitable for forecasting traffic flow at various future time periods[2].

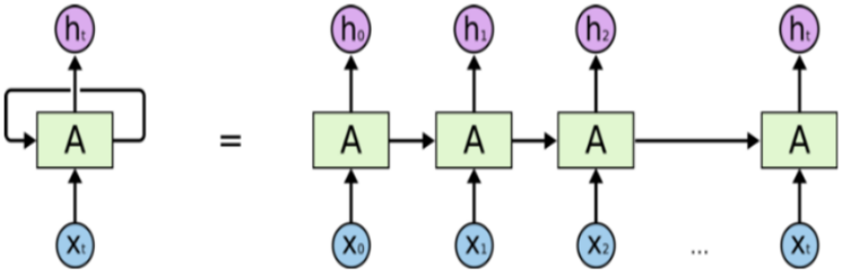

fig-2.9

The mathematical formula for a simple RNN with a single hidden layer can be expressed as follows[4]:

$$
\begin{aligned}
t_i &= \boldsymbol{W}_{hx} x_i + \boldsymbol{W}_{hh} x_{i-1} + \boldsymbol{b}_h \\
h_i &= \sigma(t_i) \\
s_i &= \boldsymbol{W}_{oh} h_i + \boldsymbol{b}_y \\
\hat{o} &= g(s_i)
\end{aligned}
$$

x_i is the input variable

$\boldsymbol{W}_{hx}$, $\boldsymbol{W}_{hh}$ and $\boldsymbol{W}_{oh}$ is weight matrixes

$\boldsymbol{b}_h$ and $\boldsymbol{b}_y$ is Bias vectors

σ and g is sigmoid function

t_i, h_i and s_i is temporary variable

$\hat{o}$ is expected Output

The cost function

$$
f = \sum_i (\left\| \hat{o}_i - o_i \right\| /2)
$$

Oi is the actual output

2.6.3 Long short-term memory (LSTM)

Long Short-Term Memory was introduced by Hochreiter and Schmidhuber (1997)[2], which is a type of recurrent neural network that was created to address the issue of vanishing gradient in traditional RNNs[2][1][5]. LSTM is mostly used when there is sequential data, like time-series forecasting, language translation, and speech recognition.

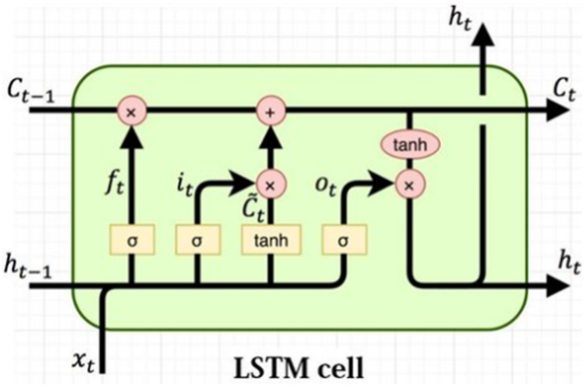

Fig-2.10 LSTM architecture

LSTM has several memory cells which can store the information for a longer period. Each cell has three gates that control the flow of information, input gate, forget gate, and output gate[5][1].

Input gate – Information added to the memory cell is controlled by the input gate. It chooses how much of the fresh input should be added to the current value of the memory cell.

Forget gate – It determines which data to keep or forget. It chooses how much of the prior cell state should be kept for the calculation of the current cell state.

Output gate – It determines how much of the memory cell should be revealed as output for the current time step.

Equation of input gate -

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$

Equation of forget gate -

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$

Equation of Output gate -

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$

Where,

$$\begin{aligned}
i_t &\rightarrow represents\ input\ gate. \\
f_t &\rightarrow represents\ forget\ gate. \\
o_t &\rightarrow represents\ output\ gate. \\
\sigma &\rightarrow represents\ sigmoid\ function. \\
w_x &\rightarrow weight\ for\ the\ respective\ gate(x)\ neurons. \\
h_{t-1} &\rightarrow output\ of\ the\ previous\ lstm\ block(at\ timestamp\ t-1). \\
x_t &\rightarrow input\ at\ current\ timestamp. \\
b_x &\rightarrow biases\ for\ the\ respective\ gates(x).
\end{aligned}$$

Further terms:

Cell state – It functions as a long-term memory and is viewed as a conveyor belt that runs

across the entire length of the network.

Candidate cell state - The network should store the new data in the cell state, and the candidate cell state is a vector that contains this data.

Equation of cell state -

$$c_t = f_t * c_{t-1} + i_t * \tilde{c_t}$$

Equation of candidate cell state -

$$\tilde{c_t} = tanh(w_c[h_{t-1}, x_t] + b_c)$$

Equation of Final output -

$$h_t = o_t * tanh(c^t)$$

Where,

$$c_t \rightarrow cell\ state(memory)\ at\ timestamp(t).$$
$$\tilde{c_t} \rightarrow represents\ candidate\ for\ cell\ state\ at\ timestamp(t).$$

The objective would be to train an LSTM network using past traffic data to forecast future traffic conditions. A time series of traffic data would constitute the network's input. The historical traffic data would be split into training, validation, and testing sets in order to train the LSTM network. Backpropagation through time (BPTT), a variation of the backpropagation technique used to train RNNs, would be used to train the network on the training set[5][2][14]. The network's hyperparameters, such as the number of hidden units and the learning rate, would be tuned using the validation set. The performance of the trained network would then be assessed using the testing set[5][2][14].

2.7 Data Analysis Methods

An important phase in the examination of traffic flow data is exploratory data analysis (EDA). The visualization, interpretation, and summarizing of data are all made possible by EDA by using various graphical and statistical techniques[8]. This method offers insights into traffic behavior, such as peak traffic hours and areas of high congestion.

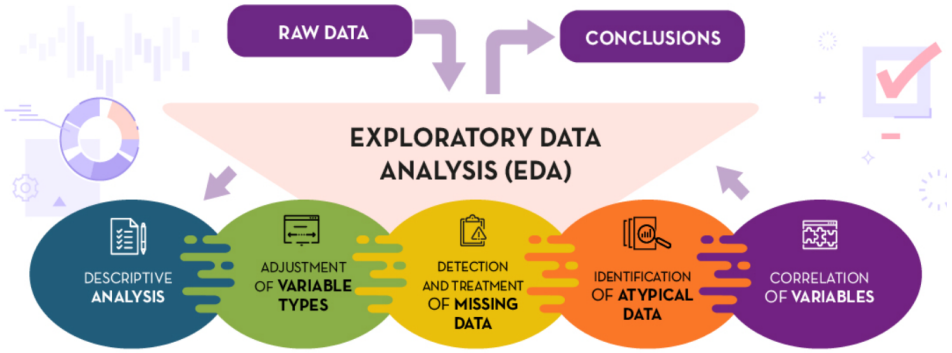

fig-2.11 EDA

EDA also aids in developing hypotheses and identifying organic patterns in the data. Techniques for feature selection are also included. EDA is a crucial stage in any research investigation since it enables the analysis of data distribution[8].

Another method used for traffic flow analysis is time series analysis. Time series analysis involves the analysis of data collected over time, such as traffic volume data collected over a period of hours, days, or weeks. This method helps in identifying trends and patterns in traffic flow, which can be used to make predictions about future traffic flow.[7]. Using this information, future traffic congestion can be predicted and solutions to lessen it can be developed. Techniques used in time series analysis include smoothing, decomposition, and forecasting. To find the underlying trend in the time series data, noise must be removed using smoothing[7].

2.8 Related works on data and models

The presence of missing data in traffic data can negatively affect traffic control and congestion prediction in intelligent traffic systems[5]. Various missing value correction methods have been proposed to address this issue. Historical imputation methods and nearest neighbor imputation methods have been suggested to provide multiple estimation values for one missing value[5].In these methods, missing data is replaced by the mean. However, these methods often produce biased results and cannot be applied when there is no data from neighboring roads. A new missing value correction method that uses past data patterns to design complete data even when there is no information from neighboring roads. The application of LSTM networks for forecasting short-term traffic flow was investigated by Zheng Zhao and others. The significance of short-term traffic forecasting in Intelligent Transportation Systems (ITS) and the difficulties encountered by conventional methodologies in handling the complexity of traffic patterns have contributed to realizing the necessity of ITS[4]. LSTM networks and deep learning, emphasize how well they can capture nonlinear temporal connections. The experimental methodology used by Zhao et al[1] includes training and testing the LSTM network using actual traffic data from a roadway in Beijing, China. The LSTM model resolves these issues. Researchers have proposed the T-LSTM model and STGCN to improve traffic flow prediction accuracy and reduce calculation time[5].

2.8.1 ARIMA

It is a traditional method that was introduced in 1970.ARIMA is a time series forecasting model that uses historical data to estimate future values. It examines data patterns and builds a mathematical model to forecast future values while taking into consideration both past values and outside factors[13]. Studies have shown that these perform well when the traffic exhibits regular variation whereas when there is an irregular variation in traffic pattern then the error is guaranteed[1].

However, it may not always be the best methodology for predicting complicated and unpredictable patterns due to the emergence of newer and more sophisticated techniques, such as deep learning models like LSTM. Despite this, ARIMA is still a widely used and beneficial method for forecasting and time series analysis[13].

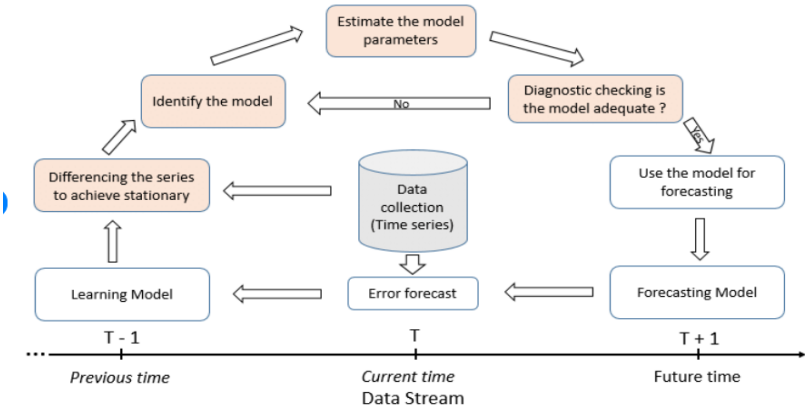

fig 2.12- Working of ARIMA

2.8.2 Simulation Models

In order to mimic traffic flow and congestion in a virtual environment, simulation models use computer software. The effectiveness of modifications to the architecture of the road network, traffic control strategies, and other interventions are frequently assessed using these models. for example SUMO, and SIMmobility[2]. One example of a simulation model is SUMO. Using these models, researchers can create virtual traffic scenarios and test out different solutions to reduce traffic congestion in a safe and controlled environment. However, these models have limitations and may not perfectly mimic real-life traffic, which can limit the accuracy of the predictions made by the simulation.

2.8.3 Deep learning models

Machine learning models are used to predict traffic congestion in the short-term and long term-by analyzing historical traffic data and identifying patterns and trends. Studies showed that there are many machine learning algorithms that are being used to predict traffic flow. some examples of machine learning models that can predict traffic congestion are the Recurrent Neural Network (RNN) Model and the Convolutional Neural Network (CNN) Model. These models can learn from traffic data over time and make predictions based on that data. They even had very good accuracy in congestion prediction, some of the models are

discussed below which were used by researchers.

2.8.4 Convolution Neural Networks

A common deep learning architecture for image classification problems is convolutional neural networks (CNNs). These are successful in classifying images because of their inherent capacity to automatically learn hierarchical characteristics from unprocessed input images[2].

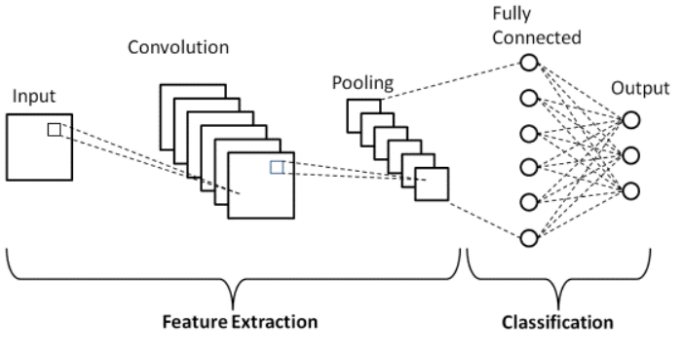

Fig-2.13 CNN architecture

A lot of researchers are using CNN methodology to predict traffic flow and they are getting pretty good results[2]. It has been used for traffic flow prediction due to its ability to extract spatial features from traffic data. CNNs use filters to find patterns in the input data, and they have been used to analyze traffic data by using inputs such as traffic flow and outputs such as traffic conditions. The input data may take the form of traffic sensor data or historical traffic flow data[11]. It has some limitations as discussed below -

1. They do not take into account temporal dependencies, which are very important when projecting traffic flow.
2. They demand a constant input size, which might be troublesome if the length of the time series data varies.

Studies have used CNN-based architectures, such as AlexNet and YOLO, to detect congestion using binary classification. These models achieve high accuracy rates, with 90.5% for AlexNet and 91.2% for YOLO, and can handle highly varied datasets[2].

2.8.5 Combination of CNN-LSTM

This hybrid approach, known as Convolutional-LSTM, uses CNNs for feature extraction and LSTMs for sequence modeling. By combining these two architectures, Conv-LSTMs can capture both spatial and temporal dependencies in the data, making them well-suited for a wide range of applications including video analysis, weather forecasting, and traffic flow prediction[2][5]. Researchers have proposed a combination of CNN and RNN models that can be used to predict traffic congestion, and CNNs can perform similarly to RNNs if given enough data[2].

The combination of the Convolutional-LSTM model has shown promising results in traffic prediction tasks, and it is widely used in research and practical applications for traffic forecasting and management.

However, it may not always be the best methodology for predicting complicated and unpredictable patterns due to the emergence of newer and more sophisticated techniques, such as deep learning models like LSTM.

3.7 Classification of Congestion

The most popular technique is to classify traffic as either "jam" or "no jam," but some researchers use a more detailed classification of "light," "medium," and "heavy" congestion[2].

3.8 Metrics Used for Evaluation

Mean Absolute Error (MAE) is a metric for evaluating how far predictions vary from actual values without taking into account whether or not the predictions were accurate. It is calculated by averaging the discrepancies between the actual value and the anticipated value while disregarding the direction of the discrepancy[2][1][5].

The root means squared error (RMSE) measures the extent to which errors exist between the values that were expected and those that were actually obtained. It is determined by averaging the squared differences between the expected and actual values and then taking the square root of that result[2][1].

Mean Squared Error (MSE) is a way to measure how well a regression model is able to predict outcomes. It works by taking the difference between the predicted value and the actual value for each data point, squaring the differences, and then taking the average of all of those squared differences[1].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}^i - y^i\right|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}^i - y^i\right)^2}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\varphi}_i - \varphi_i)^2$$

Most researchers use these metrics for the evaluation of model prediction. some of the Results of various models are displayed below

Paper	Congestion defined on the basis of:	DNN architecture	Performance	Data source	Unique aspect
(Yu et al., 2017)	Traffic speed	LSTM	MAPE: 5%	2018 VLDs (45 days) California, USA	Spatio-temporal analysis of performance
(Sun et al., 2019)	Traffic speed	CNN LSTM	90.55%≤ Accuracy≤ 96.32% 91.89%≤ Accuracy≤ 96.75%	2000 taxis GPS (28 days) Chengdu, China	Extensive sensitivity analysis w.r.t input horizon
(Cheng et al., 2018)	Traffic speed	Novel architecture↗ (built using CNN, LSTM & attention)	QWK 0.52 at 60 min	349 road links (4 months) Beijing (MapBJ)	Insights into upstream and downstream flows
(Ranjan et al., 2020)	Traffic speed	Novel PredNet↗) (built using CNN&LSTM)	Accuracy: 84.2%	Speed heat map Seoul, S Korea	Scalable architecture
(Shin et al., 2020)	Traffic speed	LSTM	MAPE: 4.29% (urban) MAPE: 6.08% (suburban)	Urban suburban areas in & around Seoul, S. korea	Observation: variation in complexity of task based on the type of network

Fig-2.14 Results of various models

3 Methodology

3.1 Flow chart

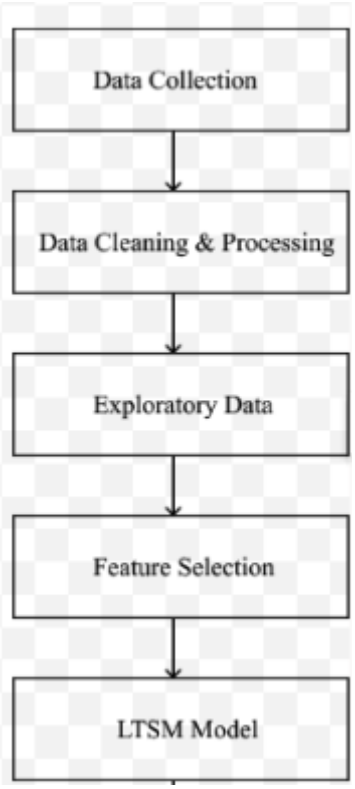

Fig-3.1 Steps

The architecture that we have followed for this project is shown above.

3.2 Data Collection

We collected raw traffic data from Transport Infrastructure Ireland(TII). TII has the data on all the road networks that are present in Ireland. So, we first had to sort the data of M50 Highway and HGVs from the raw data. The initial aim is to analyze and compare the data for 1st quarter of 2019 and 1st quarter of 2021.

3.3 Data information

The collected data consists of real-time traffic information of vehicles that pass through a series of sensors and cameras that are installed on M50 Highway. The data consist of many variables including speed, length, gap, and many more. Relevant variables are discussed below.

Cosit- It is a unique identifier for the traffic counter device. It is used to determine the location and route of the counter. It is one of the important variables which helped in sorting

out the data of M50 Highway based on the counter number.

Class/className- This indicates vehicle type, for example, car, bus, HGV, etc. It is also one of the most important variables which helped in taking out HGVs from that raw data.

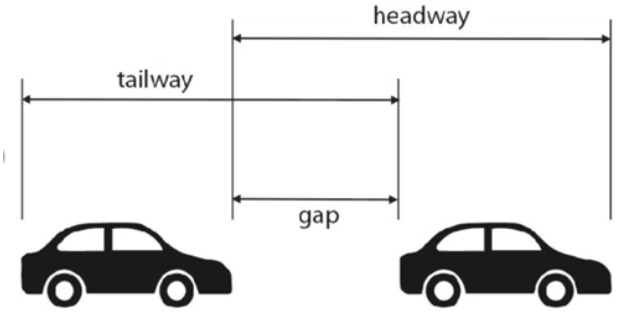

fig-3.2 Relationship between variables

Length – This is the approximate length of the vehicle recorded by the device.

Gap -This is the approximate distance between the rear of the front vehicle to the front of the vehicle behind.

Headway- This is the approximate distance between the front of the recorded vehicle and the vehicle behind. It is the sum of the length and gap.

Lane - This is the lane's unique Id, which is unique to each counter, in which the movement was recorded. It helps in determining the traffic according to the lanes.

These are the important variables that are required in analyzing the patterns of traffic on the M50 highway and further for deriving new variables which will help in the computation of Traffic.

3.4 Data cleaning and preprocessing

Data cleaning is a process in which we remove and fix all the errors in the data. Errors can be in the form of missing values, duplicate values, and incorrect values. Data preprocessing helps in transforming the data into a suitable format which is necessary for making data-driven decisions. Data cleaning and preprocessing is an important step that helps in the analysis of data accurately.

Data preprocessing involves normalization, which helps in scaling the data to a common range, or standardization. Missing values or incorrect values from the data can be replaced using a statistical method such as mean imputation. Data preprocessing helps in identifying the key features for data analysis, and by combining existing features, we are able to add new features to the data.

3.4.1 Tools and Libraries Used for Data Cleaning

1. Python language is used for the coding part. Python is one of the most popular object-oriented, high-level programming languages. As it has a wide collection of libraries(for example pandas, matplotlib, etc) that provide powerful and efficient tools

for data manipulation and analysis.

2. Jupyter Notebook is an open-source web application that enables users to create and share documents with real-time code, equations, visuals, and text. It has an interactive environment which makes it easy to apply different cleaning techniques and test the code in real time.
3. Pandas is a famous library that is commonly used for data cleaning, data manipulation, and data analysis. Pandas can significantly reduce the time and work required to prepare data for analysis. Pandas supports data manipulation, which helps in converting raw data into a more useful format. Pandas also helps in merging files but the format of data is important in order to merge. In this project, we had to merge the dataset of 1st quarter after applying the cleaning process on a si.

3.4.2 Data Cleaning Steps

First, our main aim was to remove all the unnecessary columns that were not relevant to the analysis of traffic. Below is the snippet of columns that were present in the raw data. There were 33 columns present in the raw data. Removed columns contain information that is not necessary for the analysis. Dropping them from the data frame reduces the amount of data that needs to be processed, making the analysis more efficient.

	Name	cosit	time	year	month	day	hour	minute	second	millisecond	...	speed	weight	temperature	duration	validitycode	numberofaxels	axelweights	axelspacings	unknow
0	NRA	998	2021-01-27 00:00:00.020	2021	1	27	0	0	0	20	...	70.0	0	0	0	0	0	NaN	NaN	N
1	NRA	998	2021-01-27 00:00:00.070	2021	1	27	0	0	0	70	...	71.0	0	0	0	0	0	NaN	NaN	N
2	NRA	998	2021-01-27 00:00:01.020	2021	1	27	0	0	1	20	...	71.0	0	0	0	0	0	NaN	NaN	N
3	NRA	998	2021-01-27 00:00:02.070	2021	1	27	0	0	2	70	...	70.0	0	0	0	0	0	NaN	NaN	N
4	NRA	998	2021-01-27 00:00:02.070	2021	1	27	0	0	2	70	...	69.0	0	0	0	0	0	NaN	NaN	N

5 rows × 33 columns

fig-3.3 Variable of raw data

cosit	time	year	month	day	hour	minute	second	millisecond	minutesofday	lane	lanename	class	classname	length	headway	gap	speed
1500	2021-01-27 00:00:34.000	2021	1	27	0	0	34	0	0	6	Southbound 1	6	HGV_ART	21.7	51.2	51.04	71.0
1500	2021-01-27 00:00:47.000	2021	1	27	0	0	47	0	0	6	Southbound 1	6	HGV_ART	19.4	13.3	12.20	80.0
1500	2021-01-27 00:00:53.000	2021	1	27	0	0	53	0	0	1	Northbound 1	6	HGV_ART	19.2	84.9	84.76	85.0
1500	2021-01-27 00:01:05.000	2021	1	27	0	1	5	0	1	2	Northbound 2	5	HGV_RIG	10.2	58.2	58.06	92.0
1500	2021-01-27 00:02:37.000	2021	1	27	0	2	37	0	2	6	Southbound 1	6	HGV_ART	21.1	53.9	53.70	66.0

fig-3.4 Variables after removing unnecessary columns

Second, the aim was to filter the data frame on the basis of counter number (cosit). We first looked at the counters that are present on the M50 highway. The information about the counter is present on the website in which the location of the counter and counter numbers are provided. Based on the provided data we took out counter numbers(cosit).

Third, our aim was to now filter the data on the basis of class name(type of vehicle) and remove irrelevant types of vehicles from the data frame. So now we had to filter Heavy Goods Vehicle(HGVs) and it consists of two types: HGV_RIG and HGV_ART(discussed in section 2.2) .

Fourth, filtering the data frame to keep only the rows with non-zero values for specific columns (headway, gap, speed, and length).

Headway cannot be zero because if it is zero, then it indicates that there has been an accident. So all the zeros were removed from this column.

Length of the vehicle cannot be zero because if it is zero then there is an error in the data. So all the zeros from this column were removed as well.

Gap also cannot be zero because if it is zero then it indicates that there has been an accident. So all the zeros from this column were removed too.

In this step, the code filters the DataFrame to keep only the rows with non-zero values for these columns. This is done to remove any rows with invalid data that could interfere with the analysis.

Fifth, convert specific columns in the DataFrame (cosit, lane, class, length, headway, and gap) to numeric data type using the pd.to_numeric() function.

Some columns may have been read in as strings or other data types that are not compatible with mathematical operations. In this step, the code converts these columns to numeric data type using pd.to_numeric(), which enables the data to be analyzed using mathematical operations, such as calculating the mean. The mean of values was used to fill in missing data in a dataset. This approach involves calculating the average value of the available data points for a specific variable and then using that value to replace any missing values(N/A,NaN) for that variable in the dataset.

Sixth, converting the date and time into a single column as it was given in different columns. This conversion was necessary in order to work with the date and time data more easily and to enable analysis based on specific time periods.

Seventh, the above steps are used iteratively in order to clean the data of 1st quarter of 2019 and 1st quarter of 2021. After applying the cleaning process, the data was ready for EDA which is discussed in the next section.

3.5 Exploratory Data Analysis (EDA) and Graphs

EDA and graphs are crucial tools for data analysis because they let us recognize patterns, connections, and trends in the data quickly and simply. This can assist us in making wiser decisions and effectively explaining our conclusions to others.

3.5.1 Libraries Used for Graphs Plotting

A popular Python library for data analysis and visualization is called Matplotlib. It is an effective method for analyzing and presenting traffic data since it offers a wide range of options for constructing different kinds of charts and graphs. Graphs helped in gathering many insights which are discussed in the results section.

3.5.2 Exploratory Data Analysis

It is a technique that is used to comprehend and identify a dataset's key features. In order to examine the data and find patterns, trends, and interactions between variables, statistical and visualization approaches are used[8].

EDA is an important step in understanding traffic data. It helped in examining the characteristics of the data to identify patterns, trends, and relationships between different parameters of traffic data. EDA helped in identifying congestion patterns, and peak traffic times of traffic and helped in understanding the factors that influence traffic flow. Many graphs were plotted in order to understand data, for example, hourly, weekly, and monthly vehicles on the road. The best way to understand traffic data is a lane-wise understanding of traffic. The data contains 11 lanes and my aim was to study the traffic of the busiest lane.

3.6 Use of Traffic Terminology

As we discussed traffic flow terminology(section 2.4), there are many factors that are important for the computation of traffic congestion on the M50 Highway by HGVs.

Flow = density x speed

Traffic flow, density, and speed are crucial indicators of how effectively a road or transportation network is operating. They help in locating regions that require upgrading as well as bottlenecks and locations of congestion.

Steps applied for calculation of flow density and speed -

Sampling interval = 1 Hr

Traffic data is taken of only a single lane- Northbound 2

So, first, the road length covered by vehicles in every hour is calculated, using the formula -

$$\text{len\_gap} = \Sigma(\text{Lenght+ gap}) \text{ of all the vehicles in 1 Hr}$$

As we should take summation of headway directly to calculate the road length but there is some problem in headway estimation in original data. so calculated using the above formula

1. As density is vehicles per unit length so it is calculated using (vehicle Flow in 1hr) divided by the len_gap.
2. Flow is vehicle flow per hour
3. Traffic Speed = flow/density
4. Relative speed = traffic speed/max(Traffic speed)

The result of all these are displayed in the results section

3.6 LSTM Model Creation

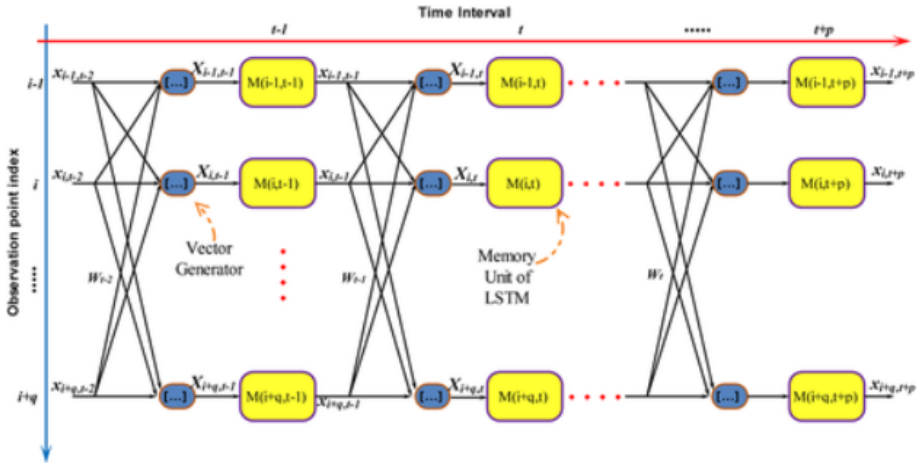

fig 3.5- 2D LSTM network

Long Short-Term memory is a type of Recurrent neural network that is used for time series forecasting[2]. The ability to assess complicated, sequential traffic data and identify long-term dependencies makes LSTM models an effective tool for traffic congestion prediction. With the help of EDA, we were able to analyze the traffic flow pattern which helped in understanding the congestion period on the M50 highway according to the lanes. The LSTM model is based on the traffic of the lanes that are present on the M50 highway.

The main reason to use LSTM is stated below:

1. Sequential nature of data
2. Long-term dependencies
3. Ability to remember previous information
4. Flexibility in input data

3.7.1 Tool And Libraries Used

The above-discussed libraries are also used in the creation of the LSTM model. Jupyter is used for the coding part of the model. In addition to this, other libraries used are discussed below -

1. TensorFlow is an open-source machine learning library that is commonly used for building deep learning models. It helps in building complex models as it provides high-level API for training neural networks, the LSTM model is defined using TensorFlow's API, and the model is trained using its optimizer and MSE loss function. It helps in applying deep learning networks and analyzing the prediction[14].
2. StandardScaler is a class from the scikit-learn library that helped in the standardization of input data and scaling it to have zero mean and unit variance. It helped in the stability and convergence of the optimization algorithm[14].

3.7.2 Model Implementation

1. The feature that I selected for the creation of the LSTM model is relative traffic speed which tells us about congestion on the road. If the value of relative speed is around 1 then there is a free flow of traffic and if the value of relative speed is near 0 then there is congestion predicted.
2. Then I defined the window_data function which helped in preparing the traffic data for the LSTM model. It makes it possible to change the data's format so that it can be used to train a neural network. The function gives the LSTM model the ability to discover temporal patterns in the input data by generating overlapping subsequences of the input data. I tried many different window sizes but got the best result with a window size of 7. A window size of 7 means the model will take Historical data of 7 days to predict the next day.
3. The next task was to split the data so that the first 80% of the input sequences and corresponding target values are used for training, and the remaining 20% are used for testing.
4. The hyperparameters described below are taken for the LSTM model to get the results:

   Batch size = 50, which determines how many windows of data are passed to the network at once during training.

   Hidden layer = 256, which determines the number of units used in the LSTM cell. This parameter regulates the model's complexity and capacity to identify long-term dependencies in the data.

   Clip margin = 4, controls the gradient clipping threshold to stop training-related gradient explosions. A gradient that exceeds this margin will have its value clipped. In this instance, a 4-clip margin is applied.

   Learning rate = 0.001, controls the step size during gradient descent optimization. it determines how quickly the model updated its weights based on the error signal.

   Epochs = 200, the number of times that the training dataset is run through the model. Each epoch corresponds to a whole iteration through the training set of data.

5. The weights and biases for all the gates: Input gate, Forget gate, Output gate, and memory cell are defined. Then define a function named LSTM_cell that computes an LSTM cell given an input, an output from a prior timestep, and the cell's current state. Using the sigmoid and Tanh activation functions, the function first calculates the input gate, forget gate, output gate, and memory cell. The input and output of the cell, along with the weights and biases.
6. The network is trained using mean squared error as the loss function and optimized using the Adam optimizer with gradient clipping.
7. A mean squared error function is used to compare the expected and observed congestion levels in order to assess the model's performance. The performance of the model improves with decreasing mean squared error.

   These steps are applied in order to create the LSTM model for Traffic congestion prediction. The results of this model are discussed in the results and evaluation section.

4 Results and Evaluation

This section consists of various graphs that are created for the analysis of traffic data for 2019 and 2021.

4.1 Traffic Fata of First Quarter of 2019 and 2021

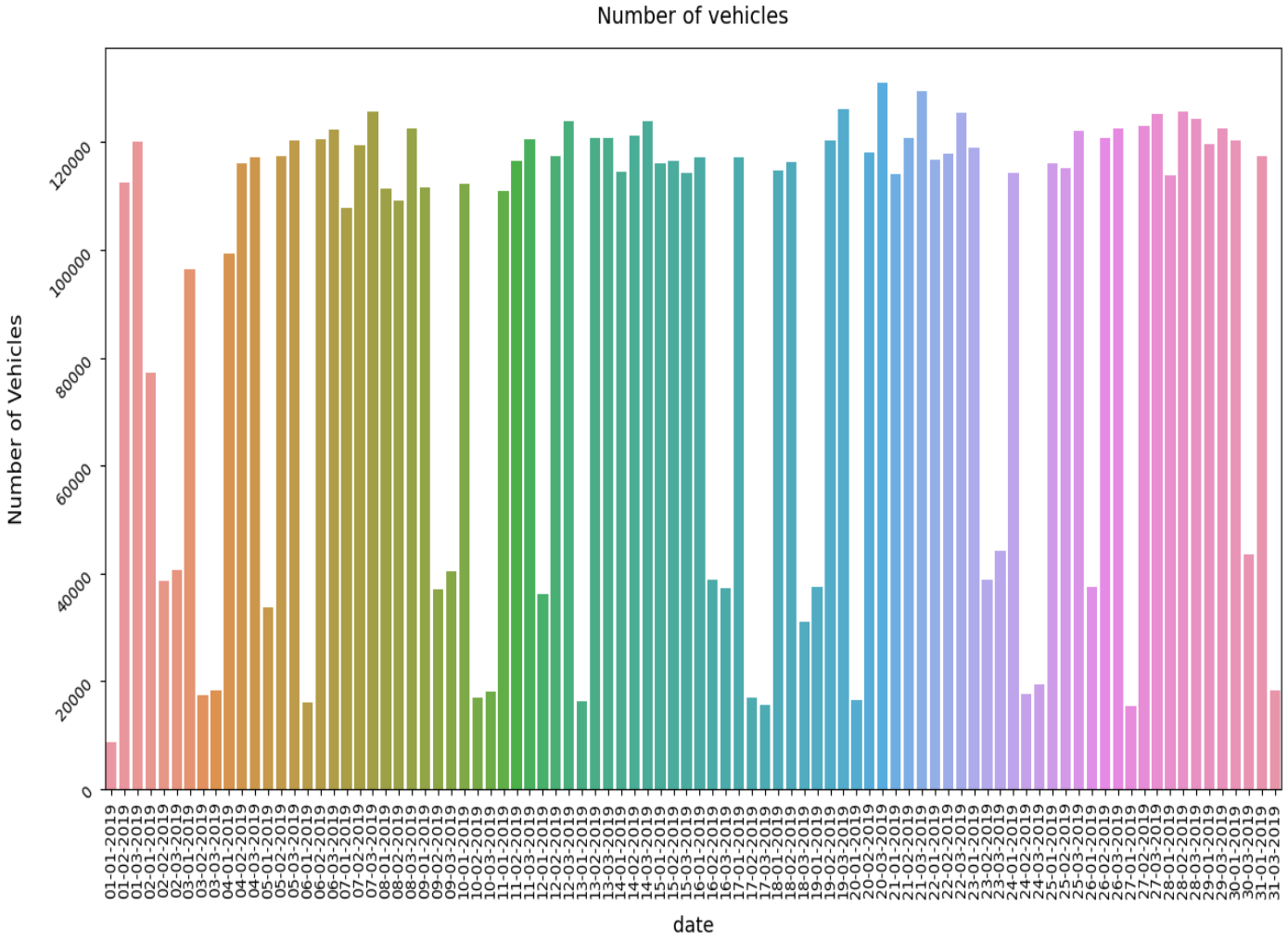

fig-4.1 Overview First Quarter of 2019

This graph depicts the number of vehicles every day passing through all the counters that are present on the M50 highway. It is in a comparative manner, for example, all the 1st Jan, Feb, and Mar are taken together. This graph gives an overview of the data for 1st quarter of 2019.

Similarly, below is the graph for 2021 which almost has a similar trend with a few exceptions which are discussed in further sections.

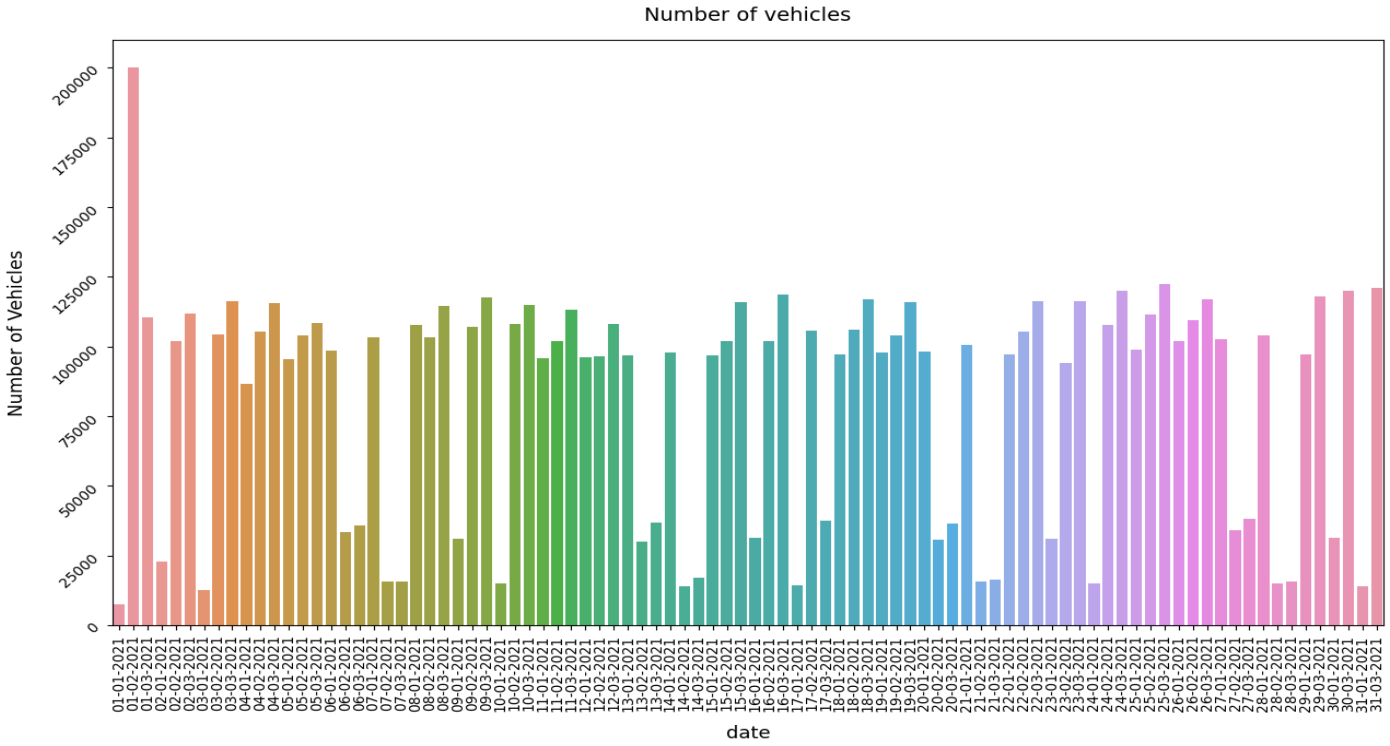

fig 4.2- Overview of 1st quarter of 2021

4.2 Week-wise distribution of Traffic

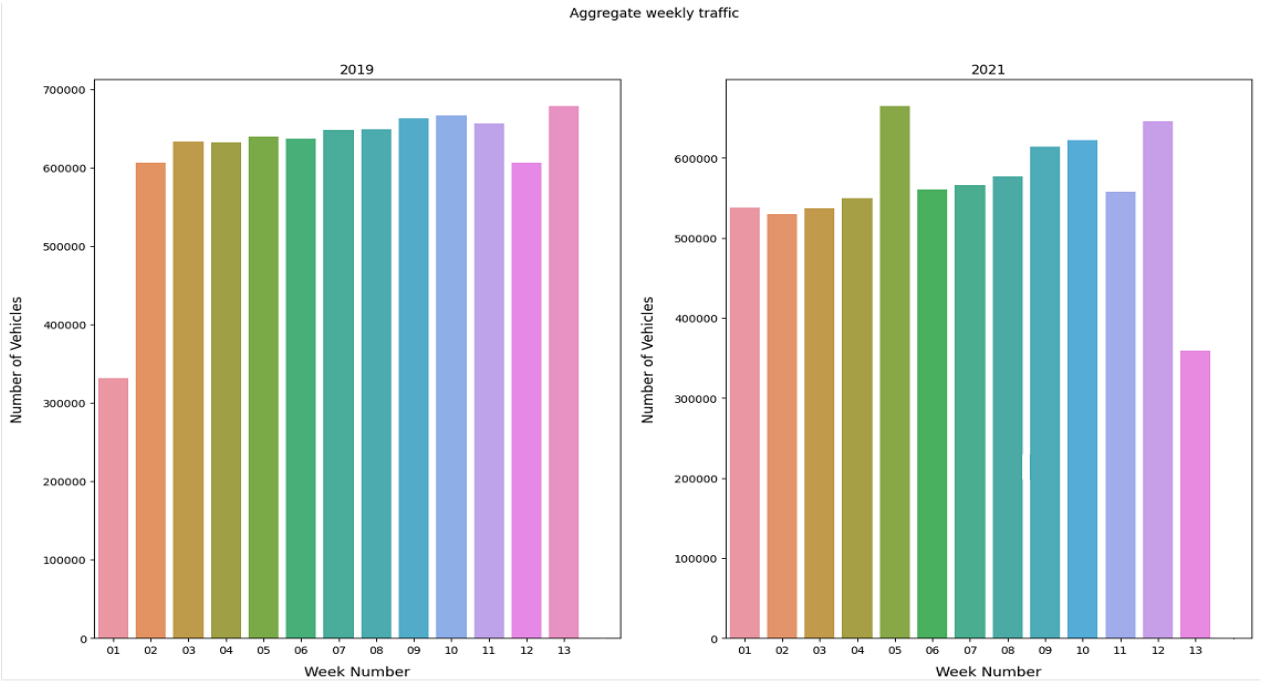

fig-4.3 Aggregate Weekly Traffic

This is the week-wise traffic of HGVs that was recorded by all the counters that are present on the M50 highway. The above graph is for the first quarter of 2019 and the first quarter of 2021. Generally, in the first week, the flow of vehicles is less because of the holidays.

The first quarter of 2019

1. In the first week of January, the number of aggregate vehicles is less and around 3,20,000 which conveys there is less traffic in the first week as compared to other weeks.
2. The pattern for the rest of the weeks goes similar which is between six lakhs to seven lakhs vehicles flowing.

The first quarter of 2021

1. There is a slight variation in traffic flow as compared to the first quarter of 2019. As we can see from the graph, in the first week of 2021, the pattern is similar to the rest of the week but in the last week, less number of vehicles are flowing.
2. As compared to 2019 there is less number of HGVs traveling in 2021. Covid-19 is one of the reasons for less number of HGVs on M50.

4.3 Aggregate Day Traffic

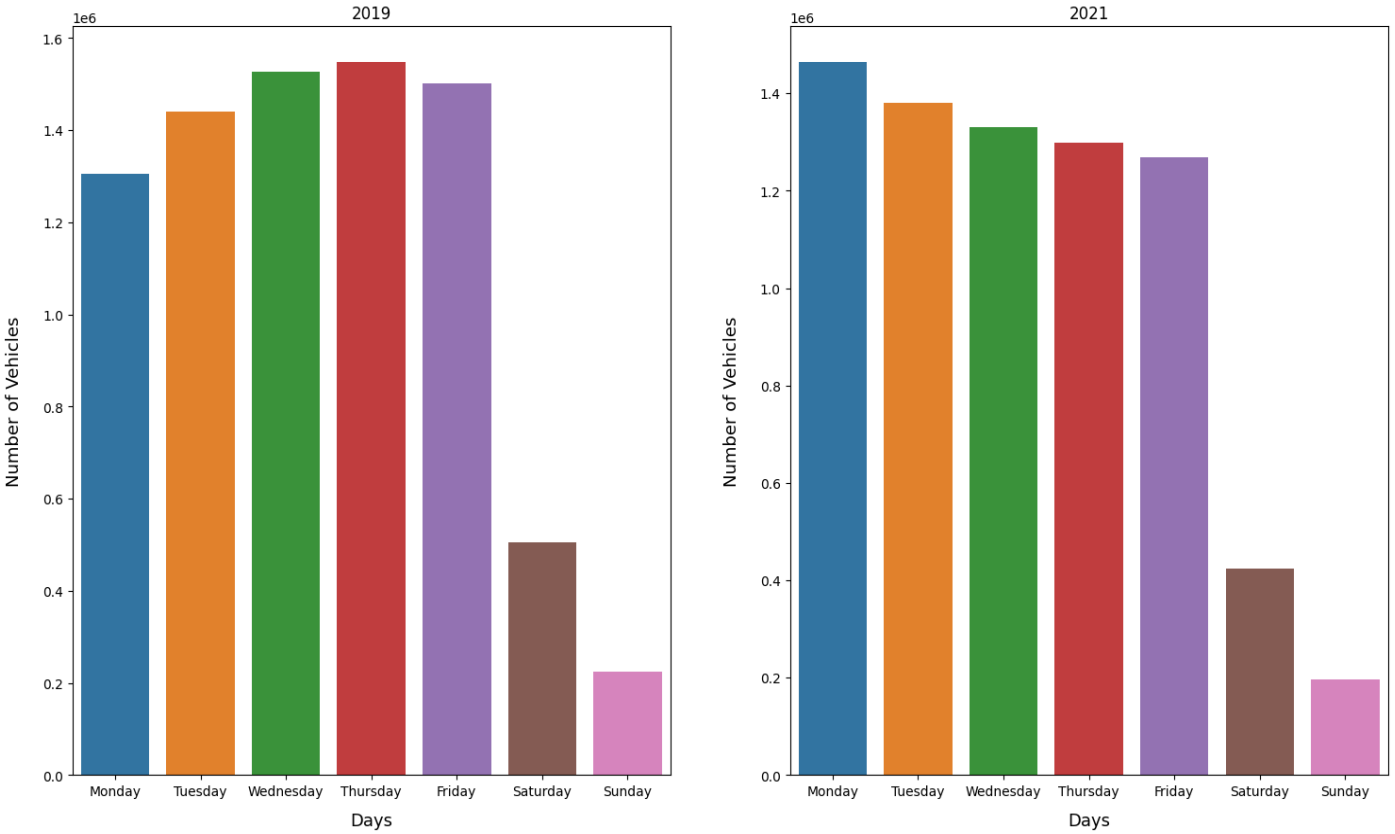

fig-4.4 Aggregate day Traffic

It is also important to analyze the data on the basis of days which tells us about which days have the highest number of vehicles. This graph gives a broad idea about which days we should focus more in order to understand the traffic patterns. As we can see from the graph, in 2019, the busiest day is Thursday whereas in 2021, Monday is the busiest.

In 2019, the flow of vehicles starts increasing after Monday and then shows a similar trend till Thursday. Whereas, in 2021, the flow of vehicles gradually starts decreasing from Monday. Weekends show a similar trend in both the years with fewer number vehicles flowing as compared to weekdays.

4.4 Aggregate Hourly Traffic

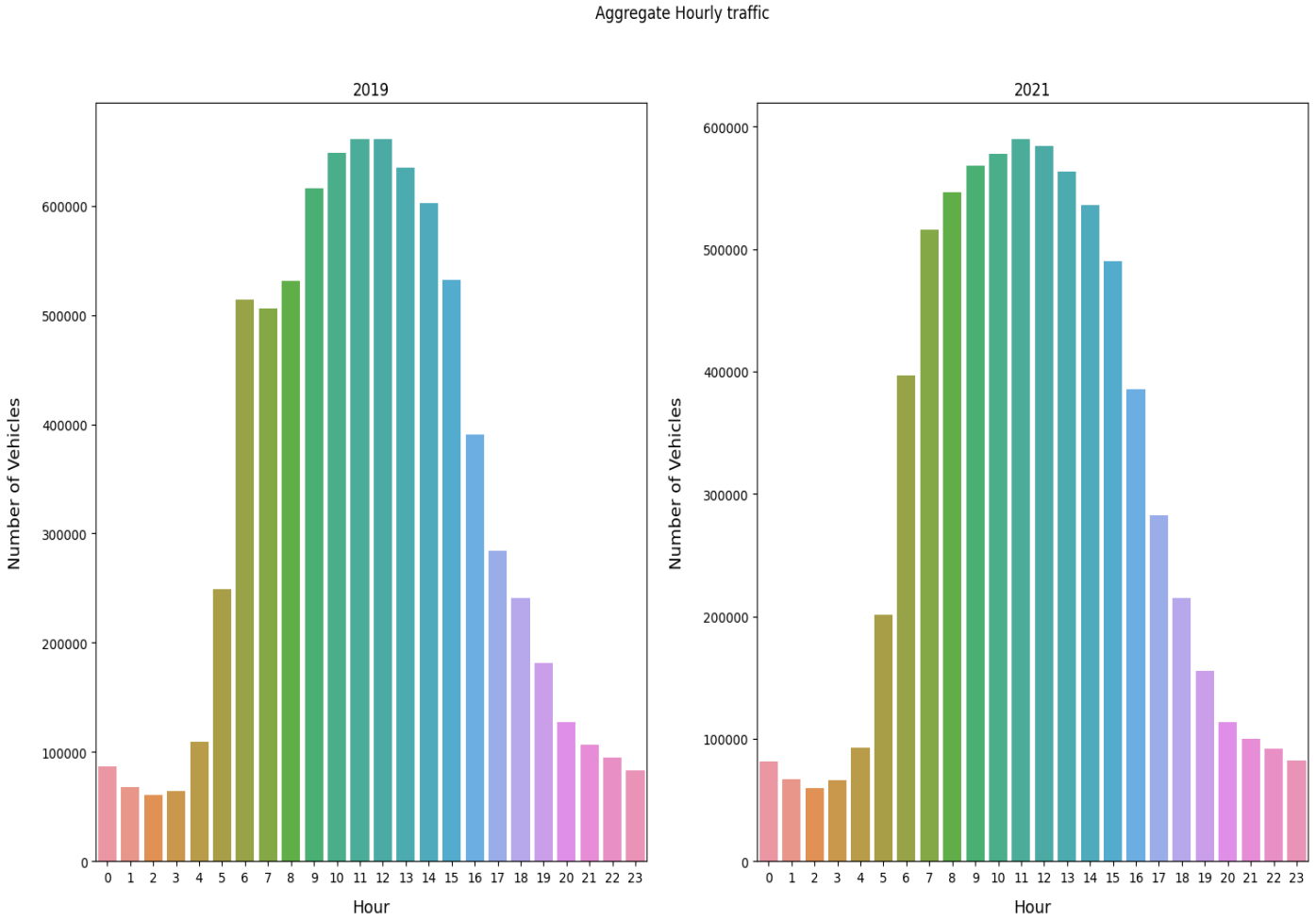

fig-4.5

This is calculated by summing up all the vehicles that are passing hourly (for example we took all the vehicles that pass from 00:00:00 to 01:00:00 on all days of the 1st quarter).

This graph is very important graph as it helps in finding out the peak traffic time. Peak time helps us in getting information about the potential traffic congestion time. According to the , initially, the flow of vehicles is low which starts increasing after 5 AM and peaks at 11 AM and 12 PM in the first quarter of both the years. The highest number of vehicles passing during peak time is around 650,000 in 2019 and around 600,000 in 2021. After this, it follows a similar pattern, that is, it gradually starts decreasing after 1 PM till 12 AM.

4.5 Lane-wise Vehicles

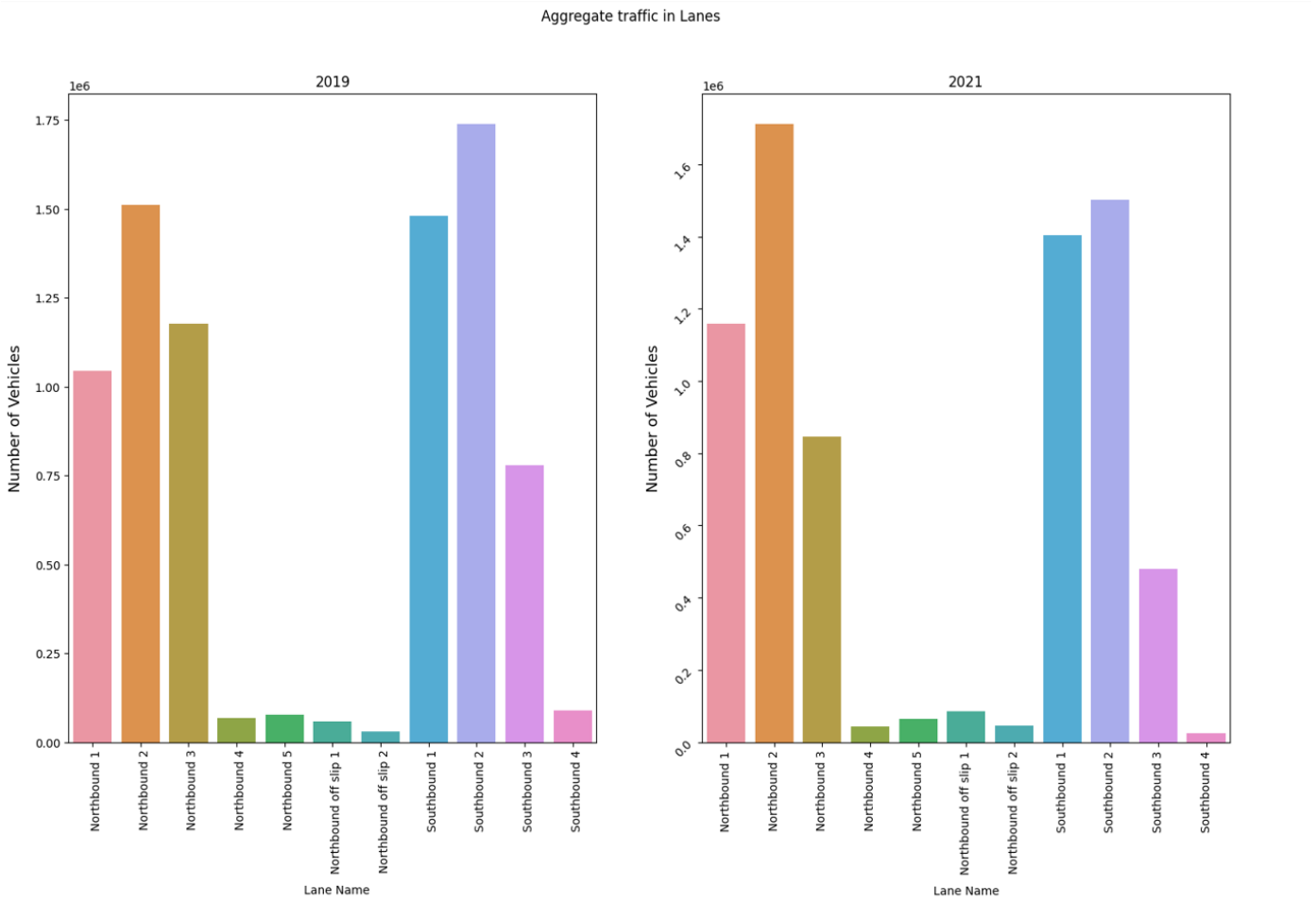

fig-4.6 Lane-wise flow of vehicles

The traffic is distributed according to the lanes. The graph shows the total number of vehicles passing(HGVs) every day on the lanes of M50. As we can see from the graph, the most busy lane is Southbound 2 and the second most busy lane is Northbound 2 in the first quarter of 2019. Whereas, in the first quarter of 2021, Northbound 2 is the busiest lane and Southbound 2 is the second most busy lane. The main aim is to focus on the busiest lanes in order to understand traffic patterns. As we can see from the graph other lanes have less traffic as compared to the busiest lane.

4.5 Analyzing weekdays and weekends

In all these graphs we are looking at the patterns of traffic on weekdays and weekends on single lane Northbound 2 as it is one of the busiest lanes. Each subgraph is made at 10min intervals after every 3 Hours for better analysis of traffic patterns.

Graphs of 21/01/2019(Monday)

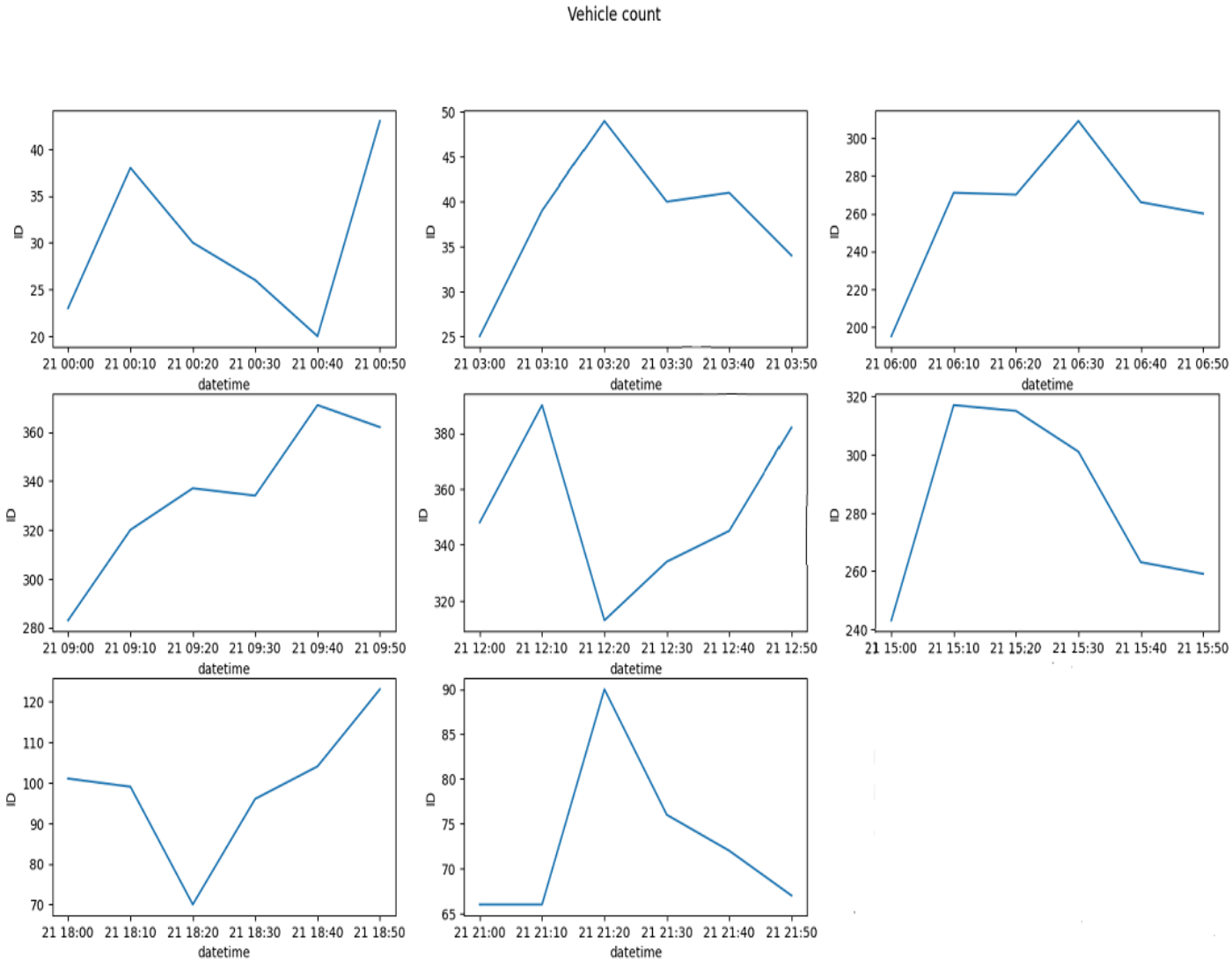

fig-4.7 graph of 10 min interval 21/01/2019

ID on the Y axis is Number of vehicles

The graphs present the traffic patterns on Northbound 2 during weekdays and weekends. The data is based on 10-minute intervals and each graph covers a 3-hour gap. The analysis of the graph for January 21, 2019, shows an increase in the number of vehicles between 9 AM and 10 AM due to office or work hours, reaching a peak of 400 vehicles around 12 PM to 1 PM, followed by a gradual decrease. The graph provides a general idea of how traffic fluctuates within a short time interval.

Graph of 18/01/2021(Monday)

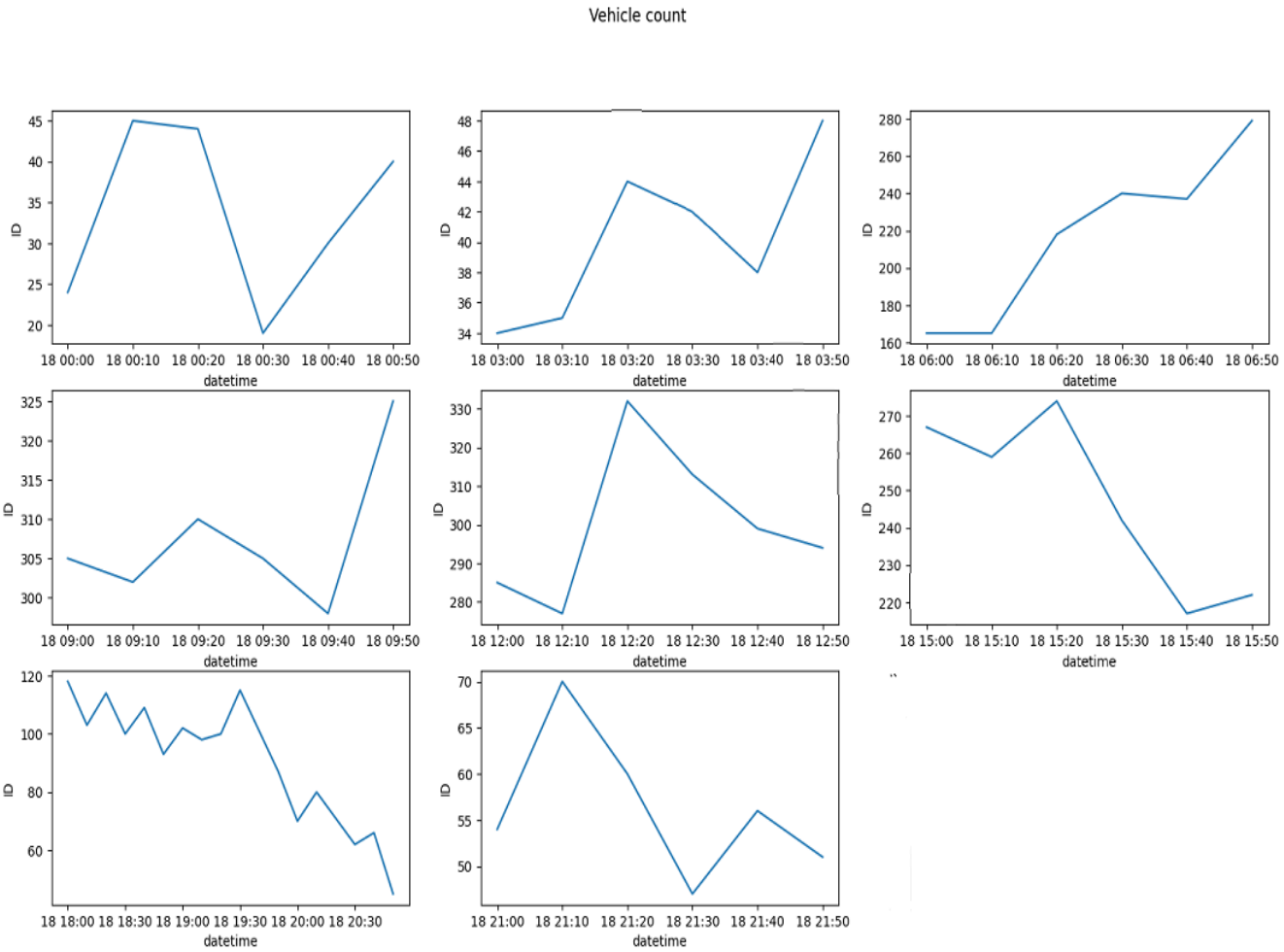

Fig 4.8 graph of 10 min interval 18/01/2021

ID on the Y axis is Number of vehicles

Similarly, the graph for January 18, 2021, depicts a rise in the number of heavy goods vehicles (HGVs) around 6 AM in the morning, while the peak traffic time is between 11 AM to 12 PM, followed by a decrease after 1 PM. The comparison of traffic patterns in 2019 and 2021 shows a similar trend.

Graphs of 26/01/2019(Saturday)

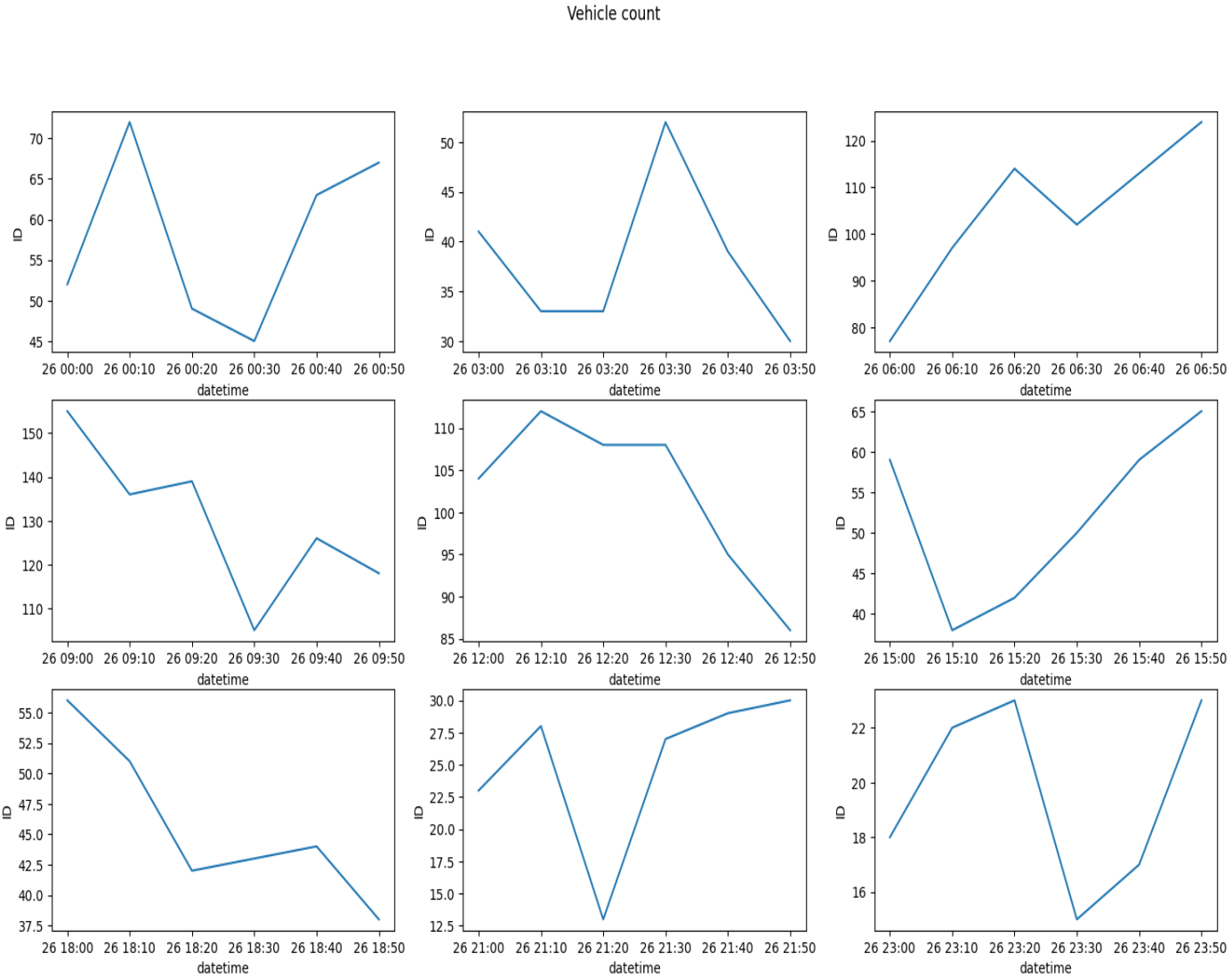

fig-4.9 graph of 10 min interval 26/01/2019

The weekend traffic graph for Northbound 2 also follows a 10-minute sampling interval and shows a different peak traffic time compared to weekdays. The peak time on weekends is between 9 AM to 20 AM while on weekdays the majority of the traffic is in the morning and afternoon, with a maximum flow between 11 AM to 1 PM, followed by a decrease after 3 PM.

Graph of 23/01/2021 (Saturday)

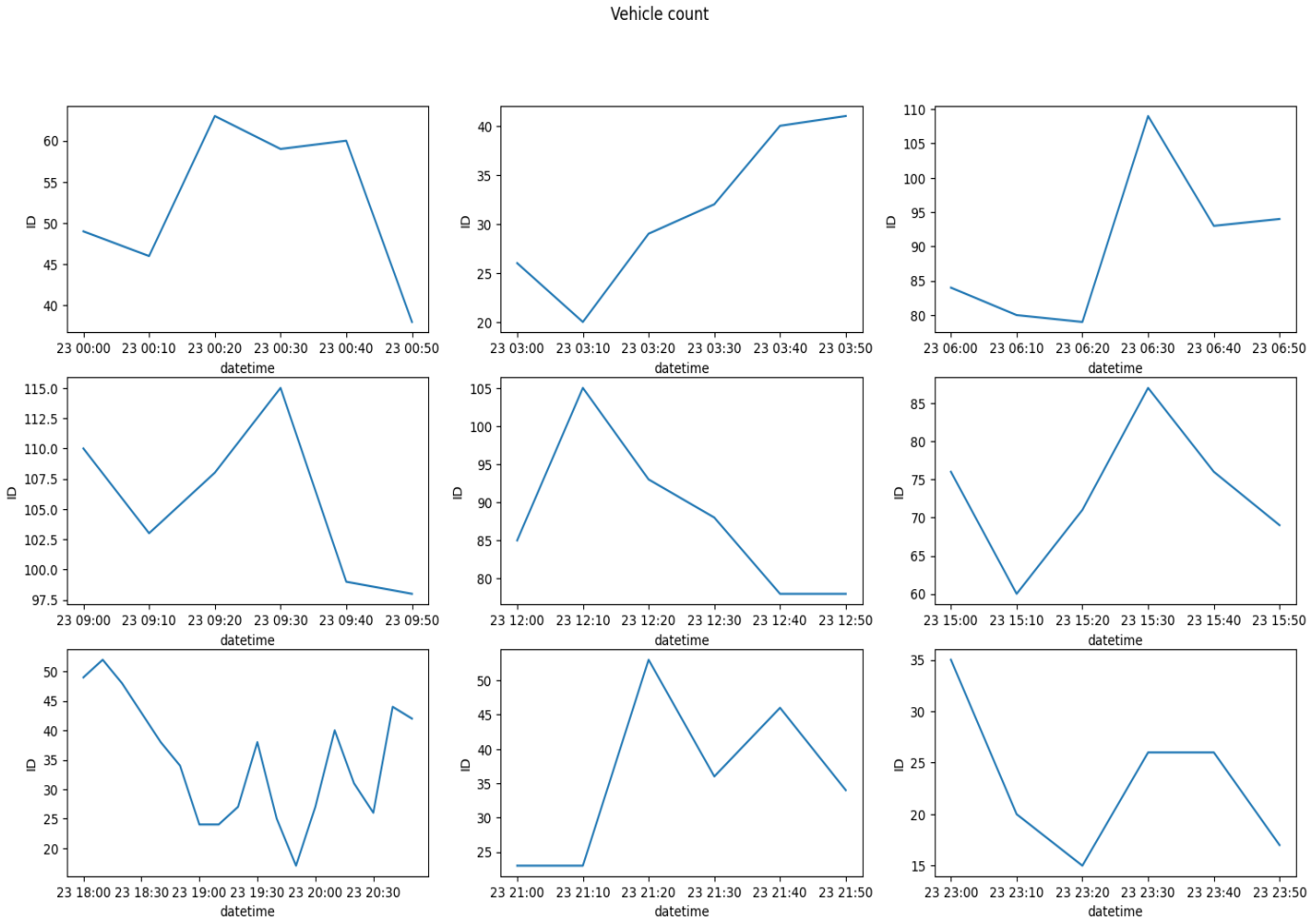

fig-4.10 -graph of 10 min interval 26/01/2019

The graph for January 23, 2021, shows a similar pattern to that of January 26, 2019, with a slight difference in the number of vehicles. In 2019, there were more HGVs than in 2021.

All the above graphs in this section provide light on the traffic patterns of the Northbound 2 lane as it is one of the busiest lanes, by showing how they change over time. For various dates in 2019 and 2021, the traffic data is examined for weekdays, weekends, and both. The graphs demonstrate that the volume of traffic follows a predictable pattern on weekdays, with peaks during morning and afternoon and decreasing numbers around lunchtime and late at night. On weekends, there is a greater fluctuation in traffic flow, with the morning rush hour witnessing the bulk of it. The figures also demonstrate that while there are minor variations in traffic volume between various dates and years, general traffic patterns are relatively constant.

5.7 Graphs of Traffic Speed, Density, Relative Speed

The data of only a single lane(Northbound 2) was analyzed as it is one of the busiest lanes. Southbound 2 has an almost similar pattern as compared to Northbound 2 but the result of Northbound 2 is more accurate and gives a clear understanding of traffic.

Data taken-1st quarter of 2019 and 2021

Lane - Northbound 2

Sampling time - 1hr which means data is divided hourly to get a better understanding

The above things are considered in all the graphs of this section.

Density

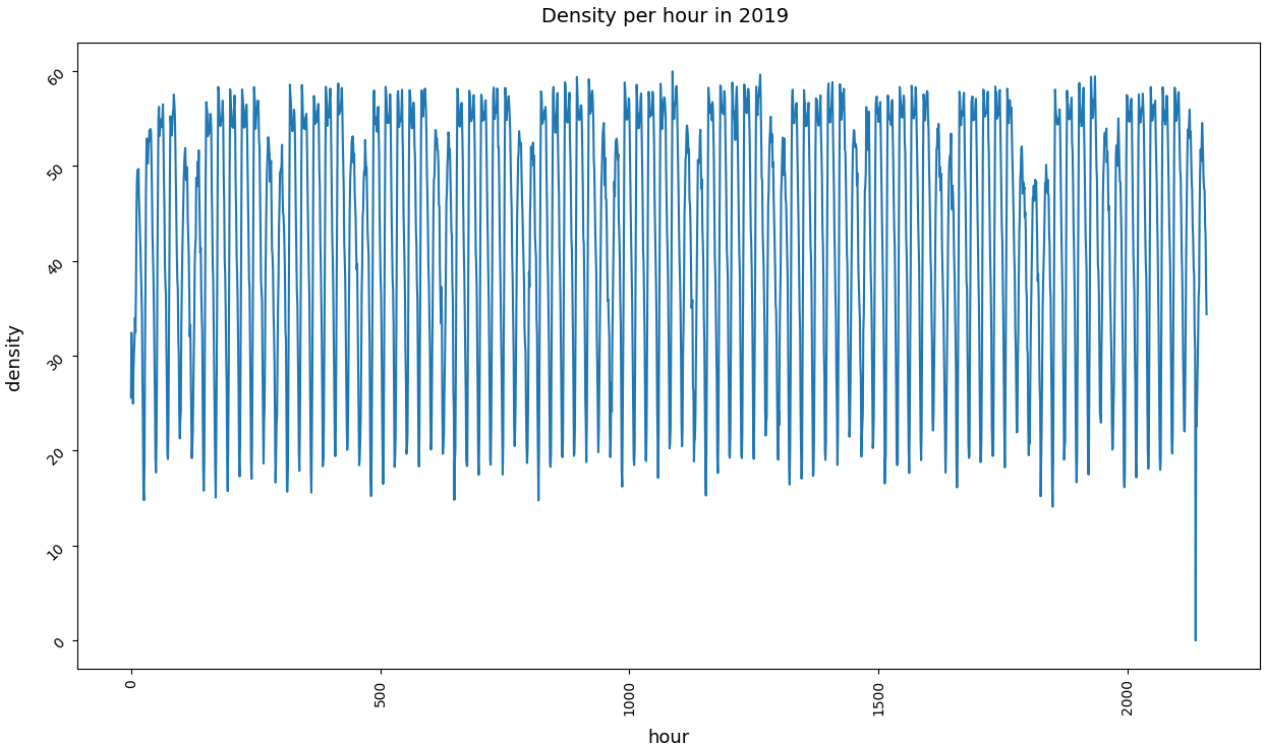

fig 4.11- density per hour in 2019

	density
count	2160.000000
mean	42.990732
std	12.769150
min	0.000000
25%	32.628810
50%	46.748859
75%	54.708971
max	59.932496

Traffic density is measured in terms of vehicles per unit length or vehicles per kilometer. As we can see from the graph, in most hours, the density is above 15 vehicles per kilometer. The calculated mean came out to be 42.9 vehicles per kilometer. 42.9 HGVs per km is regarded as reasonably high density, particularly in urban or suburban regions. Drivers may face slower speeds and longer travel times at this density due to traffic congestion and delays, which are likely to be severe. The 25th percentile density value of 32.6 vehicles per km suggests that congestion was relatively low for at least 25% of the sampling period.

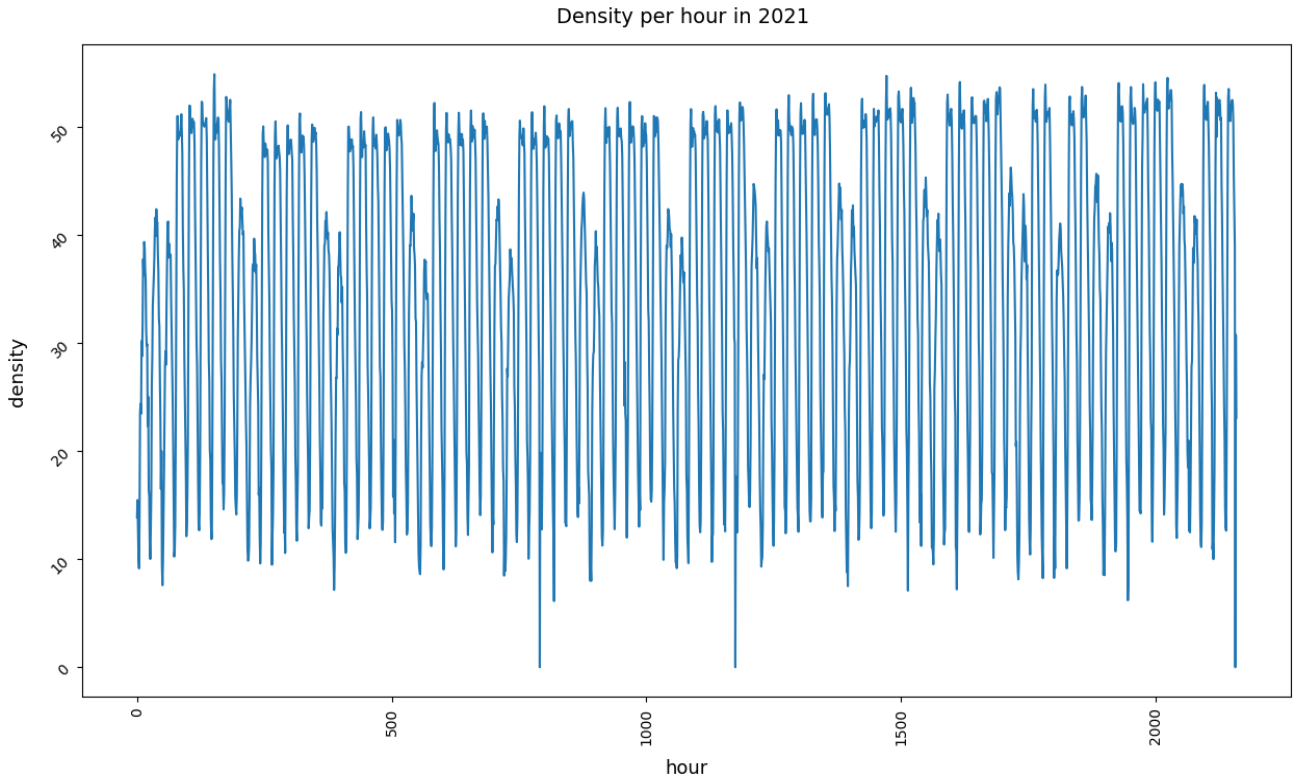

fig 4.12- Density in 2021

	density
count	2160.000000
mean	35.078962
std	13.895572
min	0.000000
25%	22.265451
50%	37.570781
75%	48.830092
max	54.918083

There were more vehicles on Northbound-2 on average in the first quarter of 2019 than there were in the same period in 2021. There are a few hours in which density is around zero whereas in 2019 there is only a single line near zero. The mean vehicle density was around 43 vehicles per km in 2019 and 35 vehicles per km in 2021. This shows that there may have been a slight reduction in congestion between the two times. The 25th percentile density value of 22.2 vehicles per km in 2021 is lower than the corresponding value of 32.6 vehicles per km in 2019. According to this, the lower end of the density distribution saw less congestion in 2021 than it did in 2019.

Traffic Speed

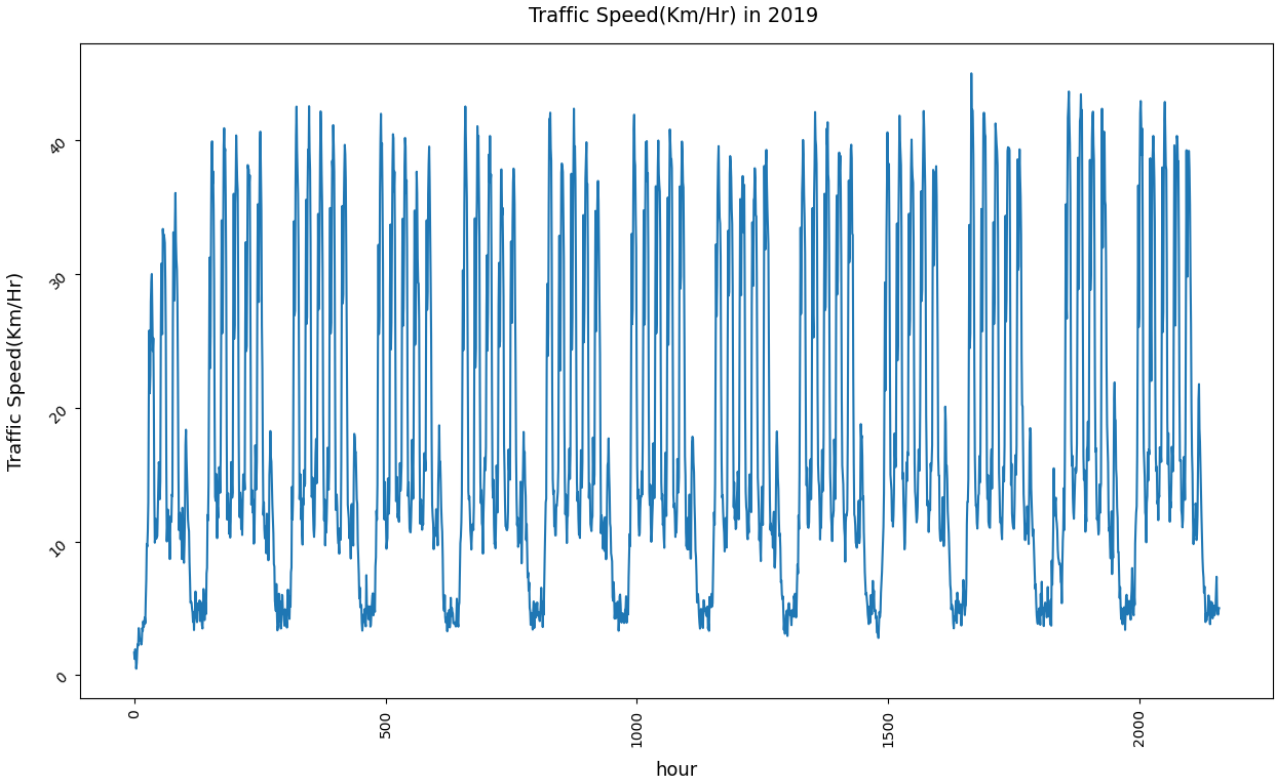

Fig 4.13 - Traffic speed per hour in 2019

	traffic_speed
count	2159.000000
mean	17.855996
std	11.694409
min	0.480940
25%	9.618885
50%	13.687910
75%	27.773785
max	45.005880

Traffic speed is the rate at which cars move along a road segment and it is expressed in terms of the distance traveled in a certain amount of time (for example, in kilometers per hour). The average traffic speed on Northbound-2 for the first three months of 2019 was 17.8 km/h, with 25% of traffic traveling at or below 9.6 km/h and 50% at or below 13.6 km/h. This means that traffic was moving at speeds of 13.6 km/hr or lower indicating that congestion was widespread.

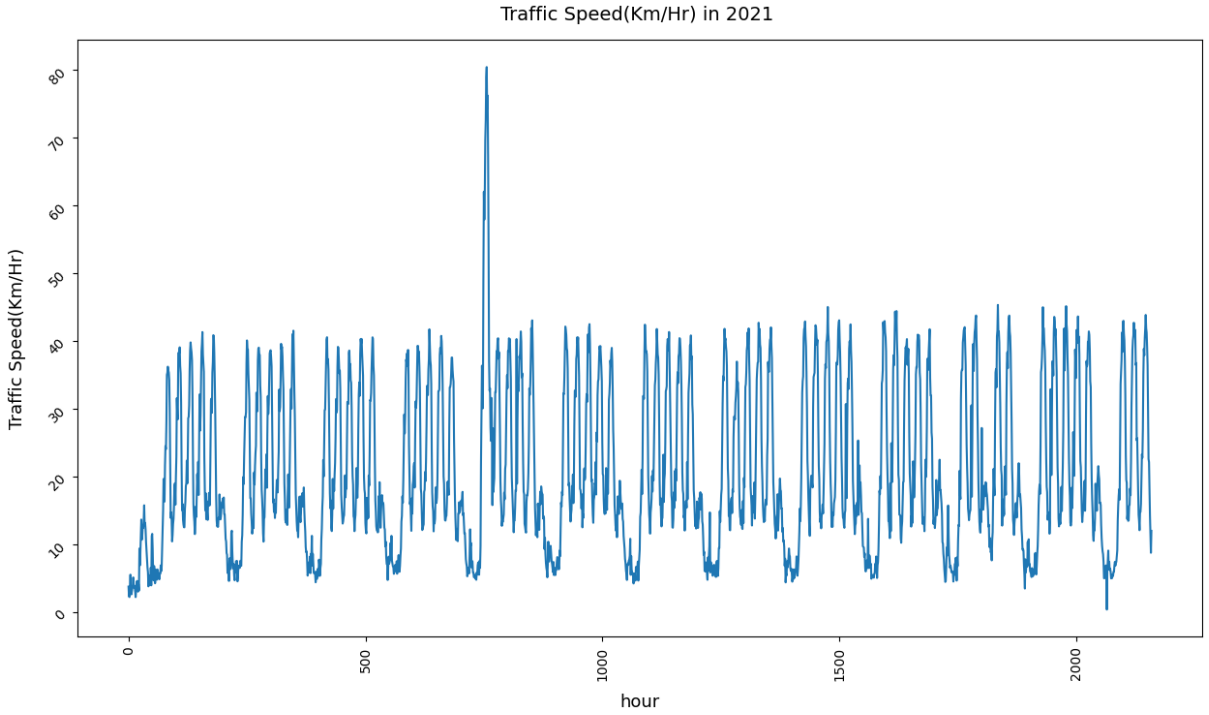

fig 4.14- traffic speed per Hour in 2021

	traffic_speed
count	2157.000000
mean	20.989237
std	12.157752
min	0.428820
25%	12.191570
50%	17.412190
75%	32.428470
max	80.449280

It appears from the characteristics of the graph that between 2019 and 2021, Northbound 2's traffic conditions have significantly improved. The mean speed of traffic in 2021 is 20.9 km/h, a significant increase over the average speed of 17.8 km/h in 2019. This means that there is less congestion and that traffic is moving more smoothly along this route. Overall the traffic speed in 2021 significantly increased as compared to 2019.

Relative Speed

Relative speed near 1 - free road

Relative speed near 0 - congestion

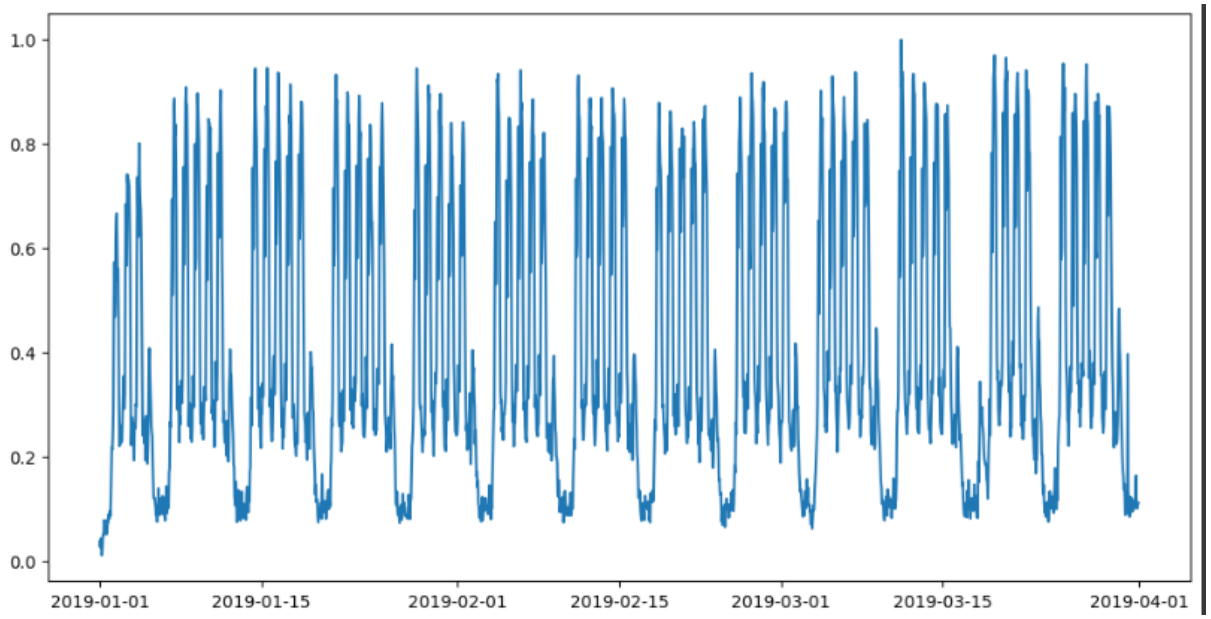

fig 4.15-relative speed per hour of 2019

Relative speed = Traffic speed/ max(Traffic speed)

count	2159.000000
mean	0.396748
std	0.259842
min	0.010686
25%	0.213725
50%	0.304136
75%	0.617115
max	1.000000

Relative speed is the difference in speed between two vehicles, such as when one vehicle is trying to pass another. The mean speed of the vehicles on Northbound 2 was only 39% of the highest speed, as the mean relative speed is 0.39. This suggests that there were significant slowdowns and delays on this route, likely due to congestion. The maximum relative speed of 1 denotes that some vehicles were capable of traveling along this route at their top speed, however, this would have been a rather uncommon occurrence.

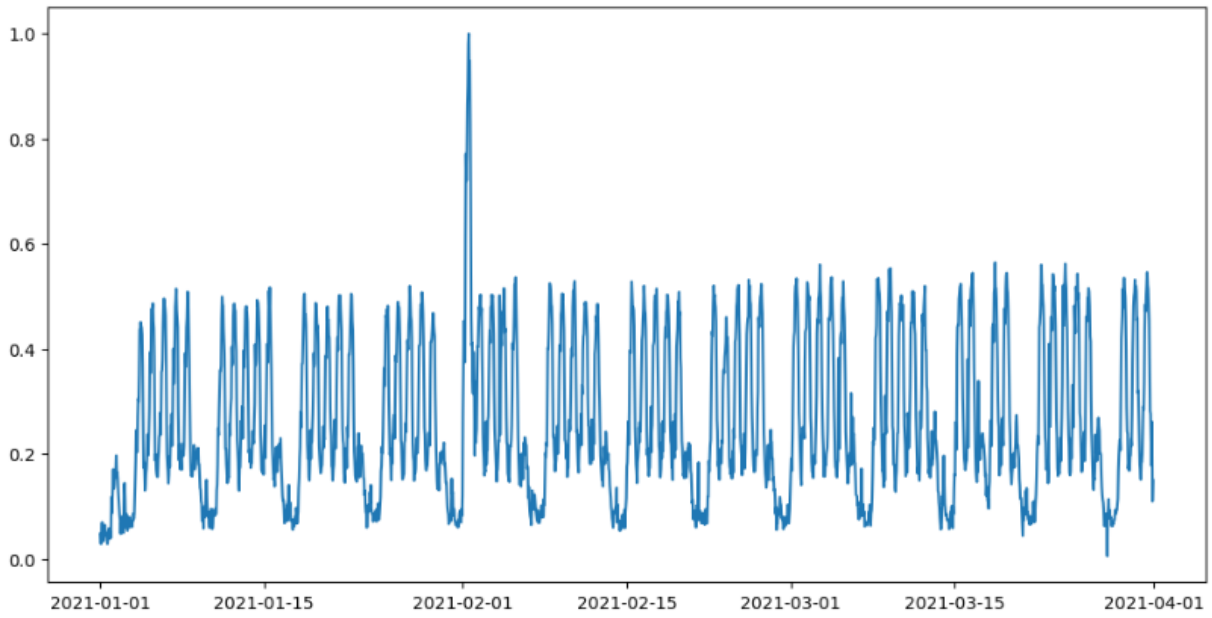

fig4.16 - Relative speed per hour of 2021

count	2157.000000
mean	0.260900
std	0.151123
min	0.005330
25%	0.151544
50%	0.216437
75%	0.403092
max	1.000000

The relative speed of traffic in 2021 is generally slow, with a mean of 0.26. The 25% of relative speeds that are less than 0.15 may indicate that congestion is a problem that frequently arises, especially in places with heavy traffic.

4.8 Graphs of LSTM trained Model

LSTM Model trained on 2019 data

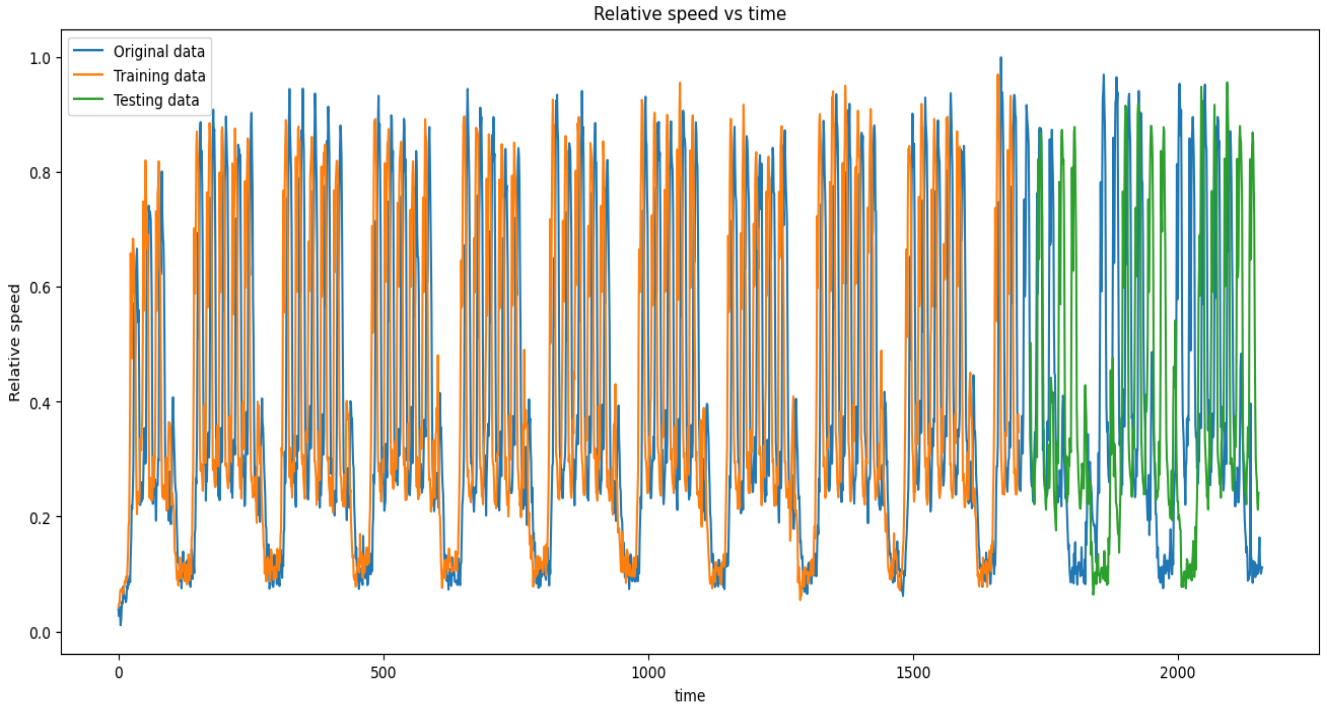

Fig-4.17 LSTM model performance on data 2019

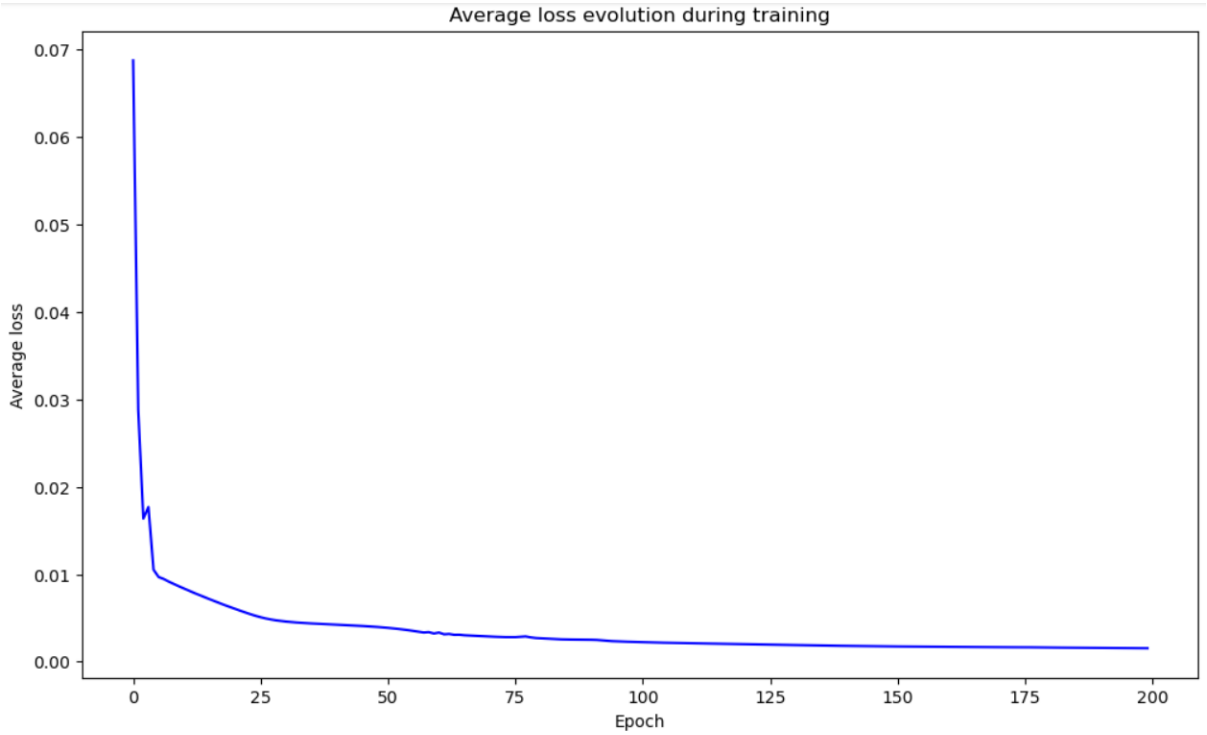

fig-4.18 MSE loss during Traning(2019)

Mean loss using MSE function **- .00153**

LSTM model trained 2021

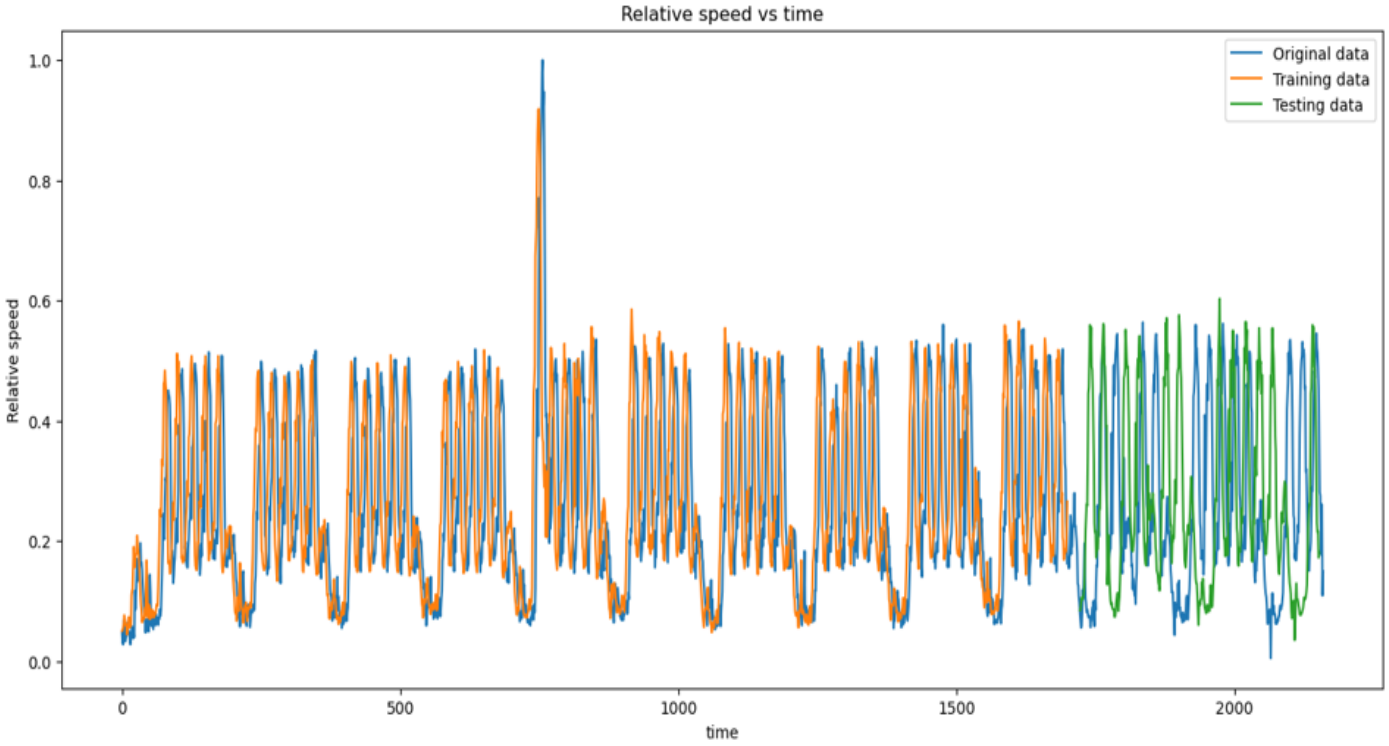

fig 4.19 -LSTM model performance on data of 2021

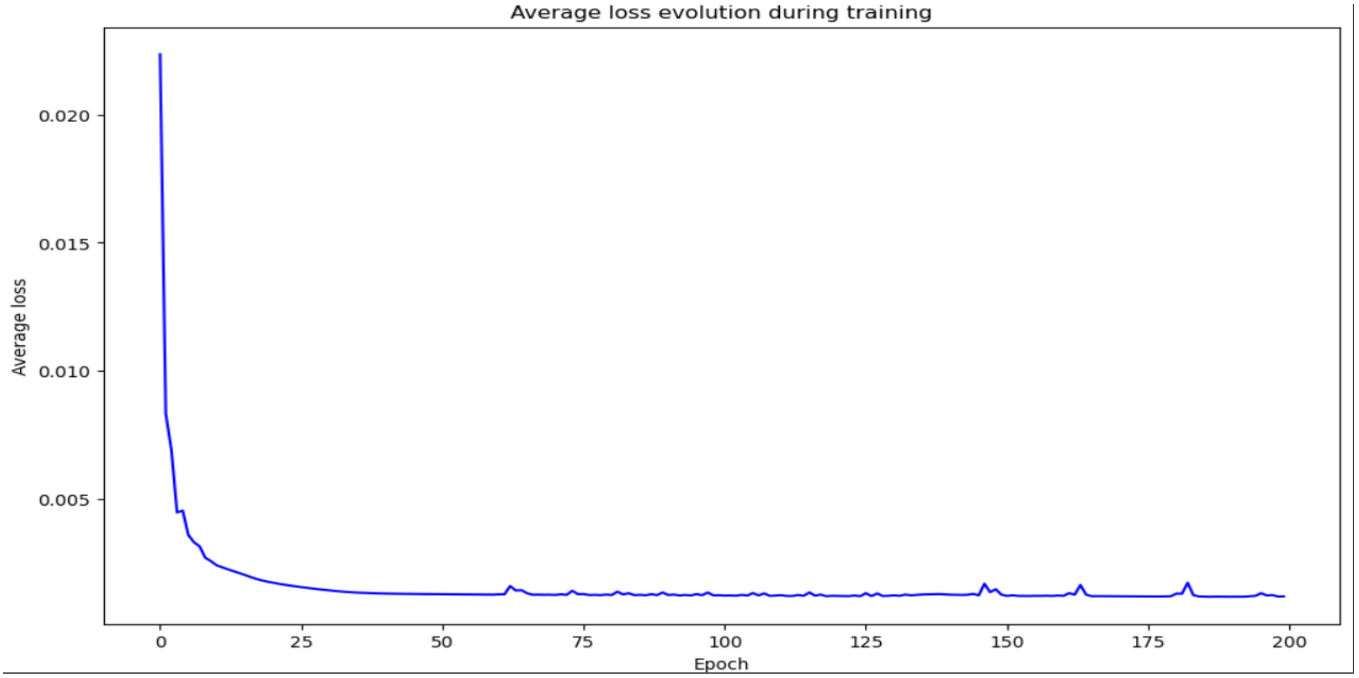

fig4.20- average loss (2021)

Mean loss using MSE function - **0.001208467**

As we can see in the graphs(fig 5.17 and fig 5.19) that the training data is overlapping with the original data it means that the model has learned from the training data and is able to

predict the original data accurately. similarly the testing data is almost overlapping with the original and follows the pattern of traffic which means the model is able to generalize well and make accurate predictions on new data. The model is able to predict the congestion by using the historical data of the past 7 days. It suggests that the model is performing well and can be used for predicting traffic with high accuracy.

The loss graph(fig 5.18 and Fig 5.20) shows the average of two-loss in every iteration. The mean loss is coming out to be **.00153** in 2019 and **.00120** in 2021 which is very small. As mean loss is coming very small which means the model has a high level of accuracy and is able to predict the values with a high degree of precision. As we increase the epoch the error decreases and further tuning its parameters by learning from the past data. As we can see in the Graph above that the predicted values are very close to the actual values and the pattern of the predicted values is almost similar to the graph of the actual values.

5 Conclusion

This research study has undertaken a comprehensive analysis of traffic flow patterns on the M50 highway in Dublin, Ireland, for the first quarter of 2019 and 2021. The graphs presented in the study helped to understand the flow of vehicles and peak traffic hours and many other factors, as well as the trends observed during weekdays and weekends. Moreover, the study calculated the mean traffic speed, mean density, and mean relative speed, which are crucial parameters used for traffic congestion analysis and prediction using the LSTM model. The LSTM model was able to learn from the training data and have high accuracy in the prediction of the original data. The model also performed well on new data, indicating good generalization ability overall, this study provides valuable insights into traffic congestion management on major highways Lane.

5.1 Limitations and difficulties

The data provided by TII (Transport Infrastructure Ireland) is a large set of data that requires preprocessing. the process can be time-consuming and requires a high-end computer that is capable of handling large amounts of data. Some of the results that are coming out from the data are not accurate as I did the analysis and found some unrealistic numbers on a few traffic parameters. One issue with the data provided by TII is that it contains incorrect values for headway as the sum of the length of each vehicle and the gap between them does not match the headway given in the data.his error has led to incorrect results and uneven graphs when analyzing the data. Another challenging task is incorporating the traffic congestion created by all the lanes on the M50 highway into a single in order to predict traffic congestion of all the Lanes.

5.2 Future Scope

The development of traffic prediction models using LSTM has shown promise in accurately forecasting traffic patterns. However, there is still room for improvement and expansion in this field, particularly in the areas of traffic rerouting and incorporating traffic from all lanes of a highway such as M50. Incorporating traffic from all lanes of a highway such as M50 is another area where LSTM models can be improved. This traffic prediction model is based on data from a single lane. Rerouting is a strategy used to alleviate traffic congestion by diverting vehicles to alternative routes that are less congested[16]. Prediction models can be used to identify potential areas of congestion in advance and suggest alternative routes to

drivers.

Bibliography

[1]LSTM network: a deep learning approach for short-term traffic forecast Zheng Zhao, Weihai Chen, Xingming Wu, Peter C. Y. Chen, Jingmeng Liu

[2]Applications of deep learning in congestion detection, prediction, and alleviation: A survey Nishant Kumar, Martin Raubal

[3] Traffic flow theory Gerlough D.L., Huber M.J.

[4]Enhancing transportation systems via deep learning: A survey Wang Y., Zhang D., Liu Y., Dai B., Lee L.H.

[5]Prediction of Traffic Congestion Based on LSTM Through Correction of Missing Temporal and Spatial Data: Dong-Hoon Shin Kyungyong Chung; Roy C. Park

[6]Comparative analysis for traffic flow forecasting models with real-life data in Beijing

Yaping Rong, Xingchen Zhang, Xuesong Feng, Tin-kin Ho, Wei Wei and Dejie Xu

[7]"An Approach to Time Series Analysis." Ann. Math. Statist. 32 (4) 951 - 989, December, 1961. https://doi.org/10.1214/aoms/1177704840

[8]Exploratory Data AnalysisMatthieu Komorowski, Dominic C. Marshall, Justin D.Salciccioli and Yves Crutain

[9]Filippo Logi, Stephen G. Ritchie,
Development and evaluation of a knowledge-based system for traffic congestion management and control,
[10]Yu R., Li Y., Shahabi C., Demiryurek U., Liu Y.
Deep learning: A generic approach for extreme condition traffic forecasting

[11]Kurniawan J., Syahra S.G., Dewa C.K., Afiahayati G.
Traffic congestion detection: Learning from CCTV monitoring images using convolutional neural network

[12]C. M. Bishop, *Neural Networks for Statistical Pattern Recognition* (Oxford University Press, Oxford, 1994)

[13]Yuhan Jia, Jianping Wu, and Yiman Du, "Traffic speed prediction using deep learning method," *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Rio de Janeiro, Brazil, 2016, pp. 1217-1222, doi: 10.1109/ITSC.2016.7795712.

[14]Real Time Traffic Flow Prediction and Intelligent Traffic Control from Remote Location for Large-Scale Heterogeneous NETWORKING USING TensorFlow S. Manikandan 1* , M. Chinnadurai 2 , D. Maria Manuel Vianny 3 , D. Sivabalaselvamani 4

[15]R. Vinayakumar, K. P. Soman and P. Poornachandran, "Applying deep learning

approaches for network traffic prediction," *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Udupi, India, 2017, pp. 2353-2358, doi: 10.1109/ICACCI.2017.8126198.

[16]A Unified Framework for Vehicle Rerouting and Traffic Light Control to Reduce Traffic Congestion Zhiguang Cao; Siwei Jiang; Jie Zhang; Hongliang Guo