

# **ANALYZING TRAFFIC DATA ON A BUSY ROAD USING R**



## **A PROJECT REPORT**

*Submitted by*

**SUGAPRIYA A(2303811724322112)**

*in partial fulfillment of requirements for the award of the course*  
**AGI1252 - FUNDAMENTALS OF DATA SCIENCE USING R**

*in*

**ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY**

(An Autonomous Institution, affiliated to Anna University Chennai and Approved by AICTE, New Delhi)

**SAMAYAPURAM – 621 112**

**JUNE- 2025**

**K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY  
(AUTONOMOUS)**

**SAMAYAPURAM – 621 112**

**BONAFIDE CERTIFICATE**

Certified that this project report on “**ANALYZING TRAFFIC DATA ON A BUSY ROAD USING R**” is the bonafide work of **SUGAPRIYA A (2303811724322112)** who carried out the project work during the academic year 2024 - 2025 under my supervision.



**SIGNATURE**

Dr.T. AVUDAIAPPAN, M.E.,Ph.D.,

**HEAD OF THE DEPARTMENT**

PROFESSOR

Department of Artificial Intelligence

K.Ramakrishnan College of Technology  
(Autonomous)

Samayapuram–621112.



**SIGNATURE**

Ms.S.Murugavalli., M.E.,(Ph.D).,

**SUPERVISOR**

ASSISTANT PROFESSOR

Department of Artificial Intelligence

K.Ramakrishnan College of Technology  
(Autonomous)

Samayapuram–621112.

Submitted for the viva-voce examination held on .....**02.06.2025**.



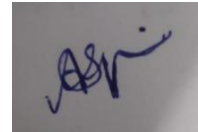
**INTERNAL EXAMINER**



**EXTERNAL EXAMINER**

## **DECLARATION**

I declare that the project report on **“ANALYZING TRAFFIC DATA ON A BUSY ROAD USING R”** is the result of original work done by me and best of my knowledge, similar work has not been submitted to **“ANNA UNIVERSITY CHENNAI”** for the requirement of Degree of **BACHELOR OF TECHNOLOGY**. This project report is submitted on the partial fulfilment of the requirement of the completion of the course **AGI1252 - FUNDAMENTALS OF DATA SCIENCE USING R**.



**Signature**

**SUGAPRIYA A**

Place: Samayapuram

Date:30.05.2025

## ACKNOWLEDGEMENT

It is with great pride that I express our gratitude and in-debt to our institution “**K.Ramakrishnan College of Technology (Autonomous)**”, for providing me with the opportunity to do this project.

I glad to credit honourable chairman **Dr. K. RAMAKRISHNAN, B.E.**, for having provided for the facilities during the course of our study in college.

I would like to express our sincere thanks to our beloved Executive Director **Dr. S. KUPPUSAMY, MBA, Ph.D.**, for forwarding to our project and offering adequate duration in completing our project.

I would like to thank **Dr. N. VASUDEVAN, M.Tech., Ph.D.**, Principal, who gave opportunity to frame the project the full satisfaction.

I whole heartily thanks to **Dr. T. AVUDAIAPPAN, M.E.,Ph.D.**, Head of the department, **ARTIFICIAL INTELLIGENCE** for providing his encourage pursuing this project.

I express my deep expression and sincere gratitude to my project supervisor **Ms.S.Murugavalli., M.E.,(Ph.D).**, Department of **ARTIFICIAL INTELLIGENCE**, for her incalculable suggestions, creativity, assistance and patience which motivated me to carry out this project.

I render my sincere thanks to Course Coordinator and other staff members for providing valuable information during the course.

I wish to express my special thanks to the officials and Lab Technicians of my departments who rendered their help during the period of the work progress.

## **INSTITUTE**

### **Vision:**

- To serve the society by offering top-notch technical education on par with global standards.

### **Mission:**

- Be a center of excellence for technical education in emerging technologies by exceeding the needs of industry and society.
- Be an institute with world class research facilities.
- Be an institute nurturing talent and enhancing competency of students to transform them as all – round personalities respecting moral and ethical values.

## **DEPARTMENT**

### **Vision:**

- To excel in education, innovation, and research in Artificial Intelligence and Data Science to fulfil industrial demands and societal expectations.

### **Mission**

- To educate future engineers with solid fundamentals, continually improving teaching methods using modern tools.
- To collaborate with industry and offer top-notch facilities in a conducive learning environment.
- To foster skilled engineers and ethical innovation in AI and Data Science for global recognition and impactful research.
- To tackle the societal challenge of producing capable professionals by instilling employability skills and human values.

## **PROGRAM EDUCATIONAL OBJECTIVES (PEO)**

- **PEO1:** Compete on a global scale for a professional career in Artificial Intelligence and Data Science.
- **PEO2:** Provide industry-specific solutions for the society with effective communication and ethics.
- **PEO3** Enhance their professional skills through research and lifelong learning initiatives.

## **PROGRAM SPECIFIC OUTCOMES (PSOs)**

- **PSO1:** Capable of finding the important factors in large datasets, simplify the data, and improve predictive model accuracy.
- **PSO2:** Capable of analyzing and providing a solution to a given real-world problem by designing an effective program.

## **PROGRAM OUTCOMES (POs)**

Engineering students will be able to:

1. **Engineering knowledge:** Apply knowledge of mathematics, natural science, computing, engineering fundamentals, and an engineering specialization to develop solutions to complex engineering problems.
2. **Problem analysis:** Identify, formulate, review research literature and analyze complex engineering problems reaching substantiated conclusions with consideration for sustainable development.
3. **Design/development of solutions:** Design creative solutions for complex engineering problems and design/develop systems/components/processes to meet identified needs with consideration for the public health and safety, whole-life cost, net zero carbon, culture, society and environment as required.
4. **Conduct investigations of complex problems:** Conduct investigations of complex engineering problems using research-based knowledge including design of experiments, modelling, analysis & interpretation of data to provide valid conclusions.
5. **Engineering Tool Usage:** Create, select and apply appropriate techniques, resources and modern engineering & IT tools, including prediction and modelling recognizing their limitations to solve complex engineering problems.
6. **The Engineer and The World:** Analyze and evaluate societal and environmental aspects while solving complex engineering problems for its impact on sustainability with reference to economy, health, safety, legal framework, culture and environment.

7. **Ethics:** Apply ethical principles and commit to professional ethics, human values, diversity and inclusion; adhere to national & international laws.
8. **Individual and Collaborative Team work:** Function effectively as an individual, and as a member or leader in diverse/multi-disciplinary teams.
9. **Communication:** Communicate effectively and inclusively within the engineering community and society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations considering cultural, language, and learning differences.
10. **Project management and finance:** Apply knowledge and understanding of engineering management principles and economic decision-making and apply these to one's own work, as a member and leader in a team, and to manage projects and in multidisciplinary environments.
11. **Life-long learning:** Recognize the need for, and have the preparation and ability for i) independent and life-long learning ii) adaptability to new and emerging technologies and iii) critical thinking in the broadest context of technological change.

## **ABSTRACT**

This project presents a comprehensive analysis of urban traffic data using the R programming language, aiming to understand traffic flow patterns, identify congestion-prone areas, and forecast peak hours. The methodology includes data collection and preprocessing, exploratory data analysis (EDA), time series forecasting using ARIMA, spatial pattern detection through K-Means clustering, and congestion classification with Random Forest. Visualizations such as line plots, boxplots, and heatmaps were used to uncover temporal and environmental factors influencing traffic volume. The results highlight key traffic trends and recurring bottlenecks, providing actionable insights for smarter traffic management and urban mobility planning. This study demonstrates the effectiveness of data-driven approaches in supporting real-time traffic solutions and long-term infrastructure development.



## ABSTRACT WITH POs AND PSOs MAPPING

### CO 5 : BUILD DATA SCIENCE USING R PROGRAMMING FOR SOLVING REAL-TIME PROBLEMS.

ABSTRACT	POs MAPPED	PSOs MAPPED
This project presents a data-driven analysis of urban traffic patterns using R programming. It aims to identify congestion-prone areas, understand peak traffic trends, and forecast future traffic flow. The workflow includes data collection, preprocessing, exploratory data analysis (EDA), time series forecasting using ARIMA, clustering with K-Means, and congestion classification using Random Forest. Various visualizations like heatmaps and line plots were used to uncover key insights. The results provide valuable inputs for smarter traffic management and long-term urban mobility planning.	<b>PO1 -3</b> <b>PO2 -3</b> <b>PO3 -3</b> <b>PO4 -3</b> <b>PO5 -3</b> <b>PO6 -3</b> <b>PO7 -3</b> <b>PO8 -3</b> <b>PO9 -3</b> <b>PO10 -3</b> <b>PO11-3</b>	<b>PSO1 -3</b> <b>PSO2 -3</b>

Note: 1- Low, 2-Medium, 3- High

## TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	<b>ABSTRACT</b>	vii
<b>1</b>	<b>INTRODUCTION</b>	1
	1.1 Objective	1
	1.2 Overview	1
	1.3 Data Science related concepts	1
<b>2</b>	<b>PROJECT METHODOLOGY</b>	3
	2.1 Proposed Work	3
	2.2 Block Diagram	3
<b>3</b>	<b>MODULE DESCRIPTION</b>	4
	3.1 Data Collection & Preprocessing	4
	3.2 Exploratory Data Analysis (EDA)	4
	3.3 Time Series Analysis & Forecasting	4
	3.4 Spatial Analysis & Heatmaps	4
<b>4</b>	<b>CONCLUSION &amp; FUTURE SCOPE</b>	6
<b>5</b>	<b>APPENDIX A SOURCE CODE</b>	7
	<b>APPENDIX B SCREENSHOTS</b>	9
	<b>REFERENCES</b>	10

# CHAPTER 1

## INTRODUCTION

### 1.1 OBJECTIVE

The primary objective of this project is to analyze traffic data using the R programming language in order to gain meaningful insights into urban traffic behavior. The study aims to monitor traffic flow patterns, identify congestion-prone areas, and predict peak traffic hours using time series models. Additionally, it focuses on visualizing traffic trends through graphs and heatmaps, detecting hotspots using clustering techniques, and classifying traffic congestion levels through supervised learning. By combining these analytical methods, the project seeks to propose data-driven solutions for better traffic management, improved signal control, and enhanced road infrastructure planning in support of smart city development.

### 1.2 OVERVIEW

This project focuses on analyzing traffic data from a busy urban road using R to uncover patterns and propose intelligent traffic management solutions. With increasing congestion in cities, understanding when, where, and why traffic builds up is essential. The project adopts a modular approach that includes data collection and preprocessing, exploratory data analysis (EDA), time series forecasting using ARIMA, spatial analysis through heatmaps, and machine learning techniques such as Random Forest and K-Means clustering. These methods help monitor traffic flow, classify congestion levels, detect hotspots, and predict future traffic trends. Ultimately, the project offers a data-driven framework to assist city planners and traffic authorities in optimizing road usage and reducing congestion.

### 1.3 DATA SCIENCE RELATED CONCEPTS

**Time Series Forecasting (for Traffic Flow Prediction): ARIMA (AutoRegressive Integrated Moving Average)**

- Best for time-series data like hourly/daily traffic volume.
- Handles trends and seasonality well.
- Suitable for historical traffic data analysis.

### **Supervised Learning (for Traffic Congestion Classification):**

#### **Random Forest (RF)**

- Works well for classifying congestion levels.
- Can handle multiple factors (time, weather, road type, etc.).

### **Unsupervised Learning (for Hotspot Detection): K-**

#### **Means Clustering**

- Groups locations with similar congestion patterns.
- Helps find recurring bottlenecks.

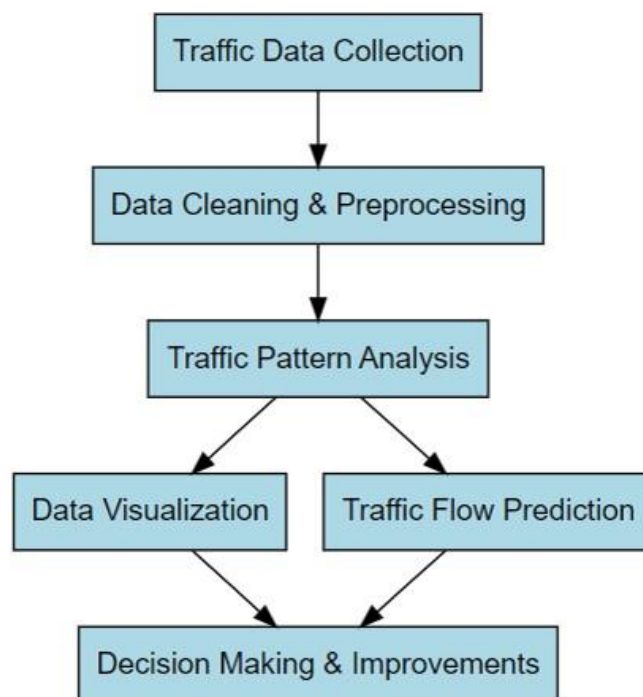
## CHAPTER 2

### PROJECT METHODOLOGY

#### 2.1 PROPOSED WORK

The proposed work aims to develop a data-driven traffic analysis system using R that can effectively monitor, predict, and manage urban traffic conditions. The system will start by collecting and preprocessing traffic data from various sources such as CSV files or sensors. Exploratory Data Analysis (EDA) will be conducted to identify trends and anomalies in traffic volume based on time, weather, and day-of-week. Time series models like ARIMA will be used to forecast traffic flow and identify peak congestion periods. Spatial patterns and congestion hotspots will be detected using K-Means clustering, while Random Forest classifiers will categorize traffic levels into low, medium, and high congestion based on features like time and weather.

#### 2.2 BLOCK DIAGRAM



# **CHAPTER 3**

## **MODULE DESCRIPTION**

### **1.1 Data Collection & Preprocessing**

- Import traffic data from CSV files or sensors.
- Handle missing values and remove anomalies.
- Convert timestamp data into usable formats (date, hour, weekday).
- Use libraries like readr, dplyr, and lubridate.

### **1.2 Exploratory Data Analysis (EDA)**

- Generate summary statistics to understand data distribution.
- Visualize traffic volume trends using histograms, boxplots, and line graphs.
- Analyze traffic behavior based on time and weather.
- Identify peak hours and fluctuation patterns.

### **1.3 Time Series Analysis & Forecasting**

- Apply ARIMA model to forecast future traffic volumes.
- Detect trends and seasonality in traffic data.
- Predict peak congestion periods for proactive traffic management.
- Use libraries like forecast and tseries.

### **1.4 Spatial Analysis & Heatmaps**

- Use K-Means clustering to group locations with similar congestion patterns.
- Identify recurring congestion hotspots.

- Create heatmaps to visualize spatial traffic intensity.
- Assist in planning infrastructure improvements and rerouting strategies.

## CHAPTER 4

### CONCLUSION & FUTURE SCOPE

This project effectively analyzed urban traffic patterns using R through a combination of data preprocessing, exploratory analysis, time series forecasting, and machine learning. Key insights revealed peak traffic during weekday rush hours, weather-dependent fluctuations, and recurring congestion hotspots identified using K-Means clustering. ARIMA models successfully forecasted traffic volume trends, while Random Forest classified congestion levels based on multiple features. The findings demonstrate the potential of data-driven approaches to enhance urban traffic management, offering valuable inputs for smarter infrastructure planning, real-time traffic control, and improved commuter experience.

#### FUTURE SCOPE

- **Real-Time Data Integration:** Extend the system to process live traffic data using APIs or IoT sensors for real-time analysis and alerts.
- **Enhanced Forecasting Models:** Implement deep learning models (e.g., LSTM) for more accurate and dynamic traffic prediction.
- **Weather & Event Impact Modeling:** Integrate external factors like accidents, festivals, or weather forecasts to improve prediction accuracy.
- **Interactive Dashboard:** Develop a Shiny or web-based dashboard for real-time monitoring, visualization, and decision support.
- **Policy and Infrastructure Planning:** Use insights from long-term data to assist city authorities in road planning, signal timing optimization, and congestion control strategies.



## APPENDICES

### APPENDIX A – SOURCE CODE

```
library(readr)
library(dplyr)
library(tidyverse)
traffic_data <- read_csv("traffic_data.csv")
traffic_data <- traffic_data %>%
filter(!is.na(Vehicle_Count)) %>%
distinct()
traffic_data$Timestamp <- as.POSIXct(traffic_data$Timestamp, format="%Y-%m-%d %H:%M:%S")
head(traffic_data)

library(ggplot2)
library(ggpubr)
ggplot(traffic_data, aes(x = Timestamp, y = Vehicle_Count)) +
  geom_line(color = "blue") +
  labs(title = "Traffic Flow Over Time", x = "Time", y = "Vehicle Count")
ggplot(traffic_data, aes(x = Vehicle_Count)) +
  geom_histogram(binwidth = 100, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Traffic Volume", x = "Vehicle Count", y =
"Frequency")

library(forecast)
library(tseries)
ts_data <- ts(traffic_data$Vehicle_Count, frequency = 24)
arima_model <- auto.arima(ts_data)
forecast_result <- forecast(arima_model, h = 48)
plot(forecast_result, main = "ARIMA Forecast for Traffic Volume")
```



```

library(prophet)
df_prophet <- traffic_data %>% select(ds = Timestamp, y = Vehicle_Count)
model_prophet <- prophet(df_prophet,
                          yearly.seasonality = TRUE,
                          weekly.seasonality = TRUE,
                          daily.seasonality = TRUE)
future <- make_future_dataframe(model_prophet, periods = 48, freq = "hour")
forecast_prophet <- predict(model_prophet, future)
plot(model_prophet, forecast_prophet)

```

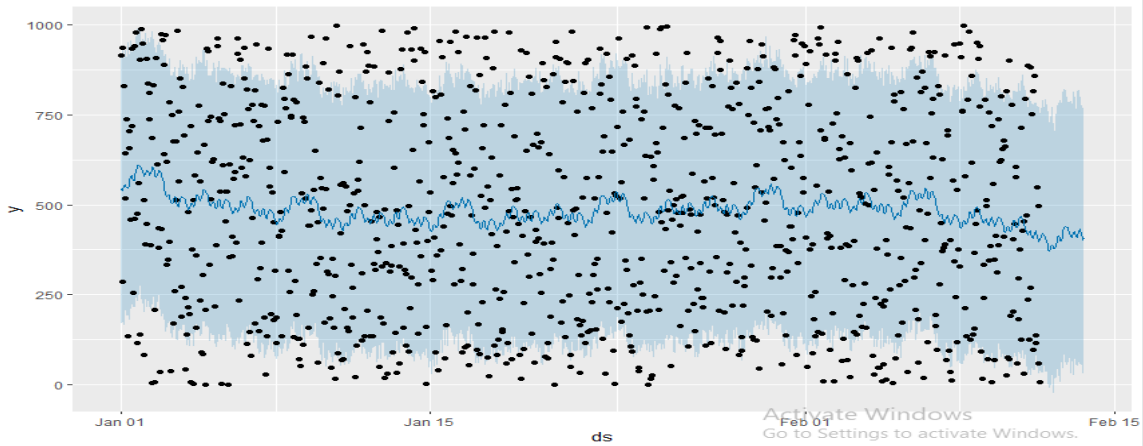
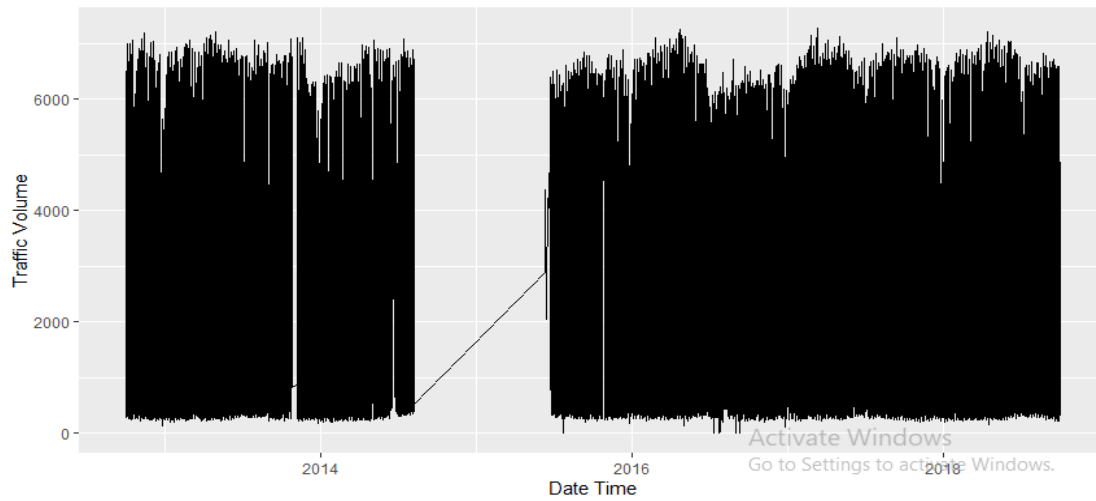
```

library(caret)
library(randomForest)
set.seed(42)
traffic_data$lat <- runif(nrow(traffic_data), min = 10.8, max = 10.9)
traffic_data$long <- runif(nrow(traffic_data), min = 78.6, max = 78.7)
traffic_data$congestion <- cut(traffic_data$Vehicle_Count,
                              breaks = c(0, 500, 1000, Inf),
                              labels = c("Low", "Medium", "High"))
table(traffic_data$congestion)
traffic_data <- traffic_data %>% filter(!is.na(congestion))
model_rf <- randomForest(as.factor(congestion) ~ Vehicle_Count + lat + long,
                        data = traffic_data)
ggplot(traffic_data, aes(x = long, y = lat, color = congestion)) +
  geom_point(size = 2) +
  scale_color_manual(values = c("green", "orange", "red")) + labs(title =
"Traffic Congestion Hotspots", x = "Longitude", y = "Latitude")

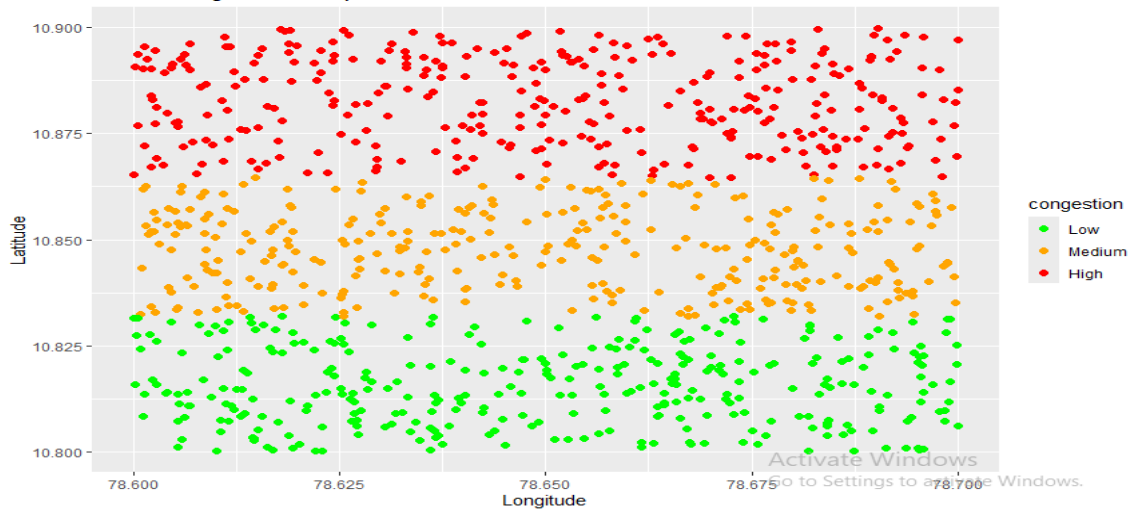
```

## APPENDIX B – SCREENSHOTS

Traffic Volume Over Time



Traffic Congestion Hotspots



## REFERENCES:

1. CRAN Task View: Machine Learning & Statistical Learning. <https://cran.r-project.org/web/views/MachineLearning.html>
2. R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>
3. R Documentation. (2024). <https://www.rdocumentation.org/>