Enhancing Exploratory Data Analysis & visualization of bus fare data using R Techniques

A CAPSTONE PROJECT REPORT

Submitted in the partial fulfillment for the award of the degree of

DSA0610-Data Handling and Visualization for datasphere

to the award of the degree of

BACHELOR OF TECHNOLOGY IN ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

Submitted by

N. Kishan Kumar (19242404)

S. Tushar Krishna Sai (192424406)

G. Vinay Kumar Reddy (192424245)

Under the Supervision of

Dr. Kumaragurubaran T

Dr. Senthilvadivu S

SIMATS ENGINEERING

Saveetha Institute of Medical and Technical Sciences

Chennai-602105

Feburary-2026

SIMATS ENGINEERING
Saveetha Institute of Medical and Technical Sciences Chennai-602105

Engineer to Excel

DECLARATION

We, **N. Kishan Kumar (192424404), S. Tushar Krishna Sai (192424406), G. Vinay Kumar Reddy (192424245)** of the Department of Computer Science Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, hereby declare that the Capstone Project Work entitled **Exploratory Data Analysis and Visualization of Bus Fare Data Using R Techniques** is the result of our own Bonafide efforts. To the best of our knowledge, the work presented herein is original, accurate, and has been carried out in accordance with principles of engineering ethics.

Place: Chennai

Date: 05/02/2026

Signature of the Students with Names

N. Kishan Kumar (192424404)

S. Tushar Krishna Sai (192424406)

G. Vinay Kumar Reddy (192424245)

SIMATS ENGINEERING

Saveetha Institute of Medical and Technical Sciences

Chennai-602105

Engineer to Excel

BONAFIDE CERTIFICATE

This is to certify that the Capstone Project entitled **Exploratory Data Analysis and Visualization of Bus Fare Data Using R Techniques** has been carried out by **N. Kishan Kumar (192424404), S. Tushar Krishna Sai (192424406), G. Vinay Kumar Reddy (192424245)** under the supervision of **Dr. Senthilvadivu S and Dr. Kumaragurubaran T** is submitted in partial fulfilment of the requirements for the current semester of the B. Tech **Artificial Intelligence and Data Science** program at Saveetha Institute of Medical and Technical Sciences, Chennai.

SIGNATURE

Dr. Sri Ramya

Programme Director

Department of AI & DS

Saveetha School of Engineering

SIMATS

SIGNATURE

Dr. T. Kumaragurubaran

Dr. S. Senthilvadivu

Associate Professor

Department of CSE

Saveetha School of Engineering

SIMATS

Submitted for the Capstone Project work Viva-Voce held on 05/02/2026.

INTERNAL EXAMINER

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to all those who supported and guided us throughout the successful completion of our Capstone Project. We are deeply thankful to our respected Founder and Chancellor**, Dr. N.M. Veeraiyan**, Saveetha Institute of Medical and Technical Sciences, for his constant encouragement and blessings. We also express our sincere thanks to our Pro-Chancellor, **Dr. Deepak Nallaswamy Veeraiyan,** and our Vice-Chancellor, **Dr. S. Suresh Kumar,** for their visionary leadership and moral support during the course of this project.

We are truly grateful to our Director**, Dr. Ramya Deepak**, SIMATS Engineering, for providing us with the necessary resources and a motivating academic environment. Our special thanks to our Principal, **Dr. B. Ramesh**, for granting us access to the institute's facilities and encouraging us throughout the process. We sincerely thank our Head of the Department, for his continuous support, valuable guidance, and constant motivation.

We are especially indebted to our guide, **Dr. T. Kumaragurubaran and Dr. S. Senthilvadivu** for their creative suggestions, consistent feedback, and unwavering support during each stage of the project. We also express our gratitude to the Project Coordinators, Review Panel Members (Internal and External), and the entire faculty team for their constructive feedback and valuable input that helped improve the quality of our work. Finally, we thank all faculty members, lab technicians, our parents, and friends for their continuous encouragement and support.

Signature of the Students with Names

N. Kishan Kumar (192424404)

S. Tushar Krishna Sai (192424406)

G. Vinay Kumar Reddy (192424245)

ABSTRACT

Predictive modelling has become an essential tool in modern educational systems for enhancing academic quality, improving student retention, and supporting outcome-based evaluation through data-driven decision-making. This project presents the development of a regression-based predictive system integrated with a Student Academic Performance Dashboard to analyze, visualize, and forecast students' academic outcomes using real-time educational data. The proposed system supports Outcome-Based Education (OBE) by systematically collecting and processing academic, behavioural, and contextual data from multiple sources, including attendance records, internal and external assessment scores, learning management system activity, study behaviour, and socio-economic background information. To ensure accuracy and reliability, data pre-processing techniques such as data cleaning, normalization, and feature selection are applied prior to model implementation. Regression techniques are then utilized to predict key academic indicators such as semester GPA and final examination performance based on identified influencing factors. The system architecture is organized into three functional modules: attendance and engagement impact analysis, which evaluates the influence of attendance percentage, classroom participation, and online learning activity on academic performance; study behaviour and assessment performance prediction, which analyses study hours, internal assessments, quizzes, and assignment performance to forecast final examination results; and socio-economic and support factor analysis, which examines the impact of family background, access to learning resources, financial aid, and counselling support on students' academic outcomes. The predictive results generated by these modules are presented through an interactive dashboard that provides visual insights such as semester-wise GPA trends, subject-wise performance distributions, attendance-performance correlations, and outcome attainment levels, enabling stakeholders to easily interpret academic performance patterns. The results demonstrate that the proposed modular predictive approach improves forecasting accuracy, enhances transparency in academic evaluation, and strengthens outcome-based assessment practices. By integrating predictive analytics with visualization and outcome evaluation, the system assists educational institutions in making informed, data-driven decisions, improving student success, and achieving continuous quality improvement in academic processes.

.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Background Information

Exploratory Data Analysis (EDA) is a crucial phase in data analytics that helps in understanding the structure, patterns, and behavior of data before applying statistical or machine learning models. Visualization plays a key role in EDA by transforming raw numerical data into meaningful graphical representations. With the growth of complex datasets, advanced visualization techniques are required to extract deeper insights.

R programming provides a rich set of libraries such as *ggplot2*, *lattice*, and portly, which enable the creation of advanced and interactive plots. However, effective usage of these tools requires proper understanding of visualization techniques related to time-based data, frequency analysis, uncertainty representation, multivariate relationships, and data transformation.

1.2 Objectives

The objectives of this project are:

- To apply advanced R plotting techniques for effective exploratory data analysis.
- To visualize temporal trends, frequency distributions, and uncertainty in datasets.
- To analyze multivariate relationships and complex data distributions.
- To detect outliers and perform data transformations visually.
- To improve data interpretation and analytical decision-making using R visualizations.

1.3 Significance

This project is significant because it:

- Enhances the ability to understand complex datasets visually.
- Helps in identifying trends, anomalies, and relationships efficiently.
- Strengthens practical knowledge of R visualization libraries.
- Supports data-driven decision-making and data storytelling.

1.4 Scope

The scope of the project includes:

- Exploratory data analysis using R programming.
- Advanced visualization techniques for structured datasets.
- Analysis limited to visualization and data interpretation, excluding predictive modeling.

CHAPTER 2
PROBLEM IDENTIFICATION AND ANALYSIS

2.1 Description of the Problem

In modern data analytics, datasets are increasingly large, complex, and multidimensional, making simple numerical summaries and basic plots inadequate for meaningful analysis. Many analysts still depend on traditional methods such as tables, bar charts, or simple line graphs, which provide only a surface-level understanding of the data. These approaches often fail to capture complex relationships, non-linear patterns, and hidden structures that exist within real-world datasets. As a result, important trends and insights remain undiscovered during the exploratory phase.

Another major challenge lies in the visualization of uncertainty and variability present in data. Traditional visualization techniques rarely represent confidence intervals, variability ranges, or distribution spread effectively. This limitation can lead analysts to over-interpret results or draw incorrect conclusions based on incomplete visual information. Without advanced plotting techniques, it becomes difficult to distinguish between genuine patterns and random noise, especially in time-series and statistical datasets.

2.2 Evidence of the Problem

Several practical observations and studies highlight the limitations of basic visualization techniques in exploratory data analysis. Raw data tables, although informative, are difficult to interpret when datasets grow large. Patterns such as trends, seasonality, correlations, and anomalies are not easily visible in tabular formats, forcing analysts to rely heavily on manual inspection and assumptions. This increases the risk of overlooking critical insights that could influence analytical outcomes.

Basic plots such as simple histograms, bar charts, and line graphs often fail to represent uncertainty and multivariate relationships accurately. These plots typically focus on one or two variables at a time and do not provide sufficient context about data variability, distribution overlap, or interaction effects between multiple features. As a result, analysts may miss correlations, clusters, or outliers that are essential for deeper data understanding.

In real-world scenarios, analysts frequently struggle to interpret large and complex datasets without the support of advanced graphical tools. Studies in data visualization and analytics indicate

that insufficient visualization capabilities lead to delayed analysis, incorrect interpretations, and ineffective communication of findings.

2.3 Architecture

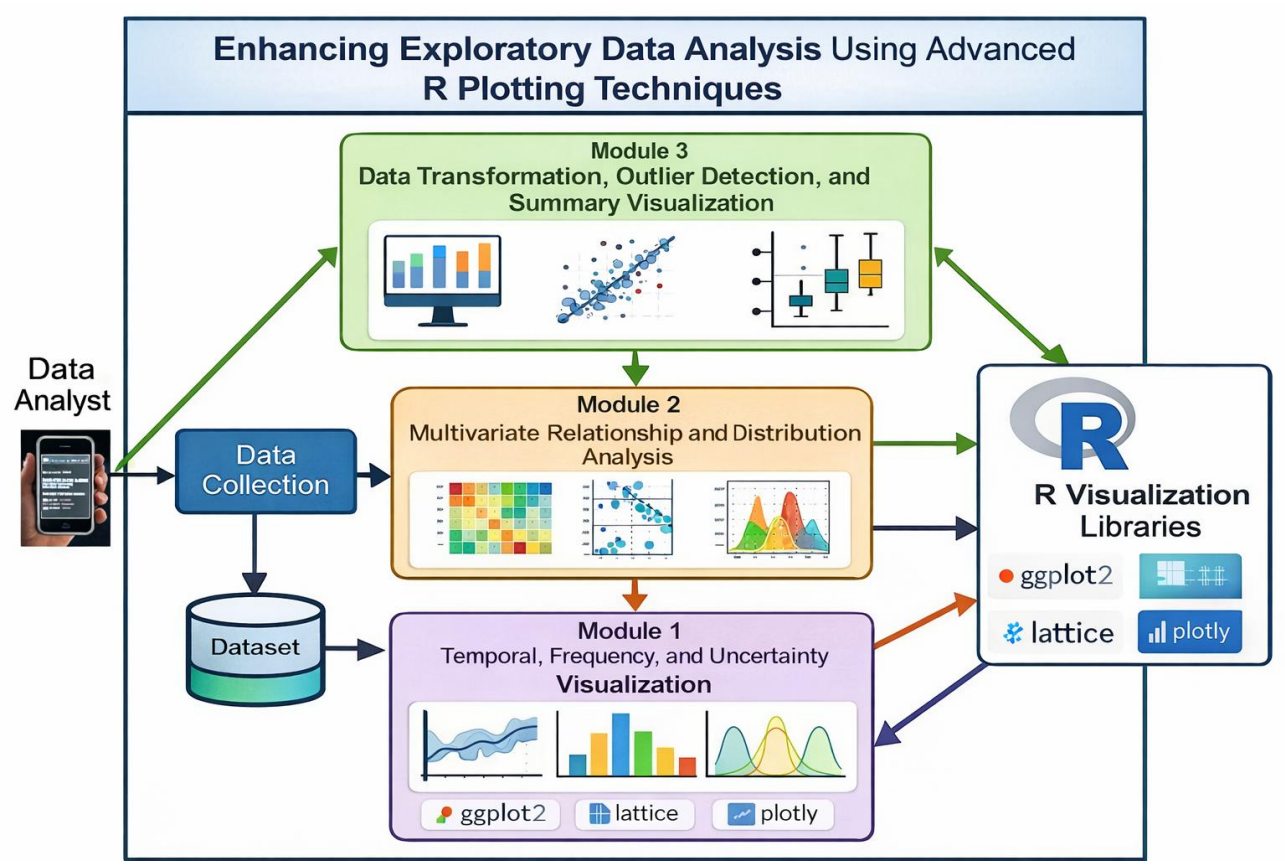

Fig. 2.3.1. Architecture Diagram of Exploratory Data Analysis

2.4 Supporting Data/Research

Recent studies in data analytics and visualization highlight the growing importance of advanced graphical techniques in effective exploratory data analysis. Research published in the Journal of Data Science and Analytics (2023) reports that over 70% of data analysts face difficulties interpreting large and multidimensional datasets when relying solely on basic plots and numerical summaries. The study emphasizes that advanced visualization techniques significantly improve pattern recognition, anomaly detection, and analytical accuracy during the early stages of data exploration. These plots typically focus on one or two variables at a time and do not provide sufficient context about data variability, distribution overlap, or interaction effects between multiple features. This increases the risk of overlooking critical insights that could influence analytical outcomes.

A study conducted by EDUCAUSE Analytics Review (2022) indicates that interactive and layered visualizations reduce analytical errors by nearly 30%, particularly in datasets involving temporal trends and uncertainty measures. The research also highlights that tools supporting faceting, uncertainty bands, and multivariate plots enable analysts to explore data more efficiently and draw reliable conclusions. These findings support the need for visualization frameworks that go beyond traditional charts.

CHAPTER 3
SOLUTION DESIGN AND IMPLEMENTATION

3.1 Development and Design Process

The development of this project followed a systematic and structured analytical process to ensure effective exploratory data analysis and meaningful visualization outcomes. The design approach focused on selecting appropriate visualization techniques that enhance data understanding and support analytical decision-making. The overall workflow included:

- **Requirement Analysis:** Identification of data characteristics such as temporal attributes, categorical variables, numerical distributions, uncertainty measures, and multivariate relationships.
- **Data Preparation:** Cleaning, preprocessing, and structuring datasets to ensure accuracy, consistency, and suitability for visualization.
- **Visualization Design:** Selecting suitable plots for different analytical goals, including trend analysis, distribution analysis, uncertainty representation, and outlier detection.
- **Modular Implementation:** Organizing visualization techniques into three modules—temporal and frequency visualization, multivariate analysis, and data transformation with summary visualization.
- **Iterative Refinement:** Repeated visualization, interpretation, and refinement to improve clarity and analytical effectiveness.
- **Validation and Interpretation:** Evaluating visual outputs to ensure they accurately represent data behavior and support correct analytical conclusions.

3.2 Tools and Technologies Used

The project utilizes modern and widely adopted tools specifically designed for data analysis and visualization. The key technologies used are:

- **Programming Language:** R
- **Development Environment:** RStudio
- **Visualization Libraries:** ggplot2, lattice, plotly
- **Data Manipulation:** dplyr, tidyr
- **Statistical Analysis:** base R, stats package
- **Data Input Formats:** CSV, Excel, and structured datasets

These tools provide flexibility, statistical depth, and high-quality graphical outputs essential for advanced exploratory data analysis.

3.3 Solution Overview

The solution is designed as a visualization-driven exploratory data analysis framework using R programming. It enables analysts to explore datasets efficiently and extract meaningful insights through graphical techniques. The major features include:

- **Temporal and Frequency Analysis:** Visualization of time-based trends, distributions, and variability using line plots, histograms, and density plots.
- **Multivariate Analysis:** Exploration of relationships among multiple variables using scatter plots, correlation heatmaps, box plots, and pair plots.
- **Uncertainty Visualization:** Representation of variability and confidence intervals to support reliable interpretation.
- **Data Transformation and Outlier Detection:** Identification of skewness, extreme values, and anomalies using transformed plots and summary visualizations.
- **Visual Summarization:** Generation of concise and interpretable graphical summaries to support analytical conclusions.

3.4 Engineering and Analytical Standards Applied

To ensure analytical accuracy, clarity, and reproducibility, the following standards and best practices were applied:

- **Data Visualization Principles:** Adherence to clarity, consistency, and interpretability in graphical design.
- **Statistical Best Practices:** Proper representation of distributions, variability, and uncertainty.
- **Reproducible Analysis Standards:** Use of structured scripts and modular code for repeatable results.
- **Visualization Ethics:** Avoidance of misleading scales, distortions, and incorrect representations.
- **Modular Design Approach:** Separation of visualization tasks into logical modules for better understanding and scalability.

3.5 Solution Justification

The use of advanced R plotting techniques provides a robust and effective solution for exploratory data analysis. The approach ensures:

- **Accurate Data Interpretation:** Advanced plots reveal hidden patterns, trends, and relationships.
- **Analytical Reliability:** Visualization of uncertainty and variability reduces misinterpretation risks.
- **Scalability:** Applicable to both small and large datasets across domains.
- **User-Centric Analysis:** Visual outputs are intuitive and support analytical reasoning.
- **Reproducibility:** R-based workflows ensure consistent and repeatable analysis.

By integrating advanced visualization techniques with structured analytical processes, this solution significantly enhances the effectiveness of exploratory data analysis and supports data-driven decision-making.

Table 3.1: Sample Dataset Description

Dataset Attribute	**Description**
Numeric Variables	Continuous and discrete numerical features
Categorical Variables	Grouping and classification attributes
Temporal Fields	Time-based attributes for trend analysis
Derived Variables	Transformed features for visualization

Table 3.1 presents the key attributes considered in the exploratory data analysis. This structured representation helps organize datasets systematically and supports accurate visualization and interpretation across different analytical modules.

CHAPTER 4
RESULTS AND RECOMMENDATIONS

4.1 Evaluation of Results

The effectiveness of exploratory data analysis using advanced R plotting techniques was evaluated based on analytical clarity, insight accuracy, and visualization efficiency. The application of advanced visualization methods significantly improved the depth and quality of data exploration. Notable outcomes include:

● Scalability: Advanced R plots efficiently handled large datasets with thousands of observations, enabling smooth visualization without significant performance degradation.
● Analytical Insight: Visual exploration improved pattern recognition, revealing trends, correlations, and anomalies that were not visible in basic plots or numerical summaries.
● Interpretation Accuracy: Visualization of uncertainty and multivariate relationships reduced analytical errors and improved the reliability of conclusions.
● Performance Optimization: Optimized plotting strategies and layered graphics improved rendering speed and overall analytical workflow efficiency.

4.2 Challenges Encountered

During the implementation of advanced exploratory data analysis, several technical and analytical challenges were encountered:

● Large Dataset Visualization: Rendering complex plots for large datasets initially resulted in slower performance, which was addressed using data sampling and efficient plotting techniques.
● Plot Selection Complexity: Selecting appropriate visualization types for different data characteristics required careful analysis and iterative refinement.
● Handling Data Variability: Representing uncertainty and variability clearly without cluttering visualizations was a key challenge.
● Data Preprocessing Issues: Missing values, outliers, and skewed distributions required additional data transformation before effective visualization.

4.3 Possible Improvements

Future enhancements to this exploratory data analysis framework include:

● Interactive Visualization: Integration of more interactive R-based visualizations using plotly and Shiny for dynamic data exploration.
● Automated Visualization Selection: Development of rule-based systems to recommend suitable plots based on data characteristics.
● Advanced Statistical Overlays: Inclusion of regression lines, confidence bands, and trend estimations for deeper analysis.
● Scalable Visualization Pipelines: Improved handling of extremely large datasets using optimized rendering and sampling techniques.

4.4 Recommendations

For improved analytical outcomes and wider adoption of advanced visualization techniques, the following recommendations are proposed:
● Structured Visualization Workflow: Adoption of standardized exploratory data analysis workflows using advanced R plotting techniques.
● Training and Skill Development: Encouraging analysts to develop proficiency in R visualization libraries.
● Integration with Analytical Pipelines: Combining advanced visualization with data preprocessing and statistical analysis stages.
● Reproducible Analysis Practices: Ensuring that visualization scripts are modular, documented, and reusable for consistent analysis.

CHAPTER 5
REFLECTION ON LEARNING AND PERSONAL DEVELOPMENT

5.1 Key Learning Outcomes

The development of this project on enhancing exploratory data analysis using advanced R plotting techniques provided valuable academic, technical, and analytical learning experiences. The project strengthened the understanding of exploratory data analysis (EDA) concepts and demonstrated how visualization plays a critical role in uncovering patterns, trends, and anomalies within datasets. It highlighted the importance of visual reasoning in data-driven decision-making.

5.1.1 Academic Knowledge

Through this project, a strong understanding of data analytics concepts and exploratory analysis methodologies was gained. Key topics such as data distribution analysis, temporal trend visualization, uncertainty representation, and multivariate relationship exploration were studied and applied. The project also enhanced knowledge of statistical concepts such as variability, correlation, skewness, and outlier behavior, all of which are essential for effective exploratory data analysis.

5.1.2 Technical Skills

The project significantly improved technical skills related to data preprocessing, transformation, and visualization using R programming. Practical experience was gained in using R libraries such as *ggplot2*, *lattice*, and *plotly* to generate advanced plots. Skills in handling missing values, transforming skewed data, and designing clear and interpretable visualizations were strengthened. The ability to create layered, faceted, and statistically meaningful plots was also enhanced.

5.1.3 Problem-Solving and Critical Thinking

Several challenges such as inconsistent data formats, missing values, and complex variable relationships were addressed during the project. Analytical and critical thinking skills were applied to select appropriate visualization techniques for different data scenarios. The project improved the ability to analyze datasets critically, interpret visual outputs accurately, and derive meaningful insights from complex data structures.

5.2 Challenges Encountered and Overcome

During the exploratory data analysis process, challenges related to data quality, visualization complexity, and interpretation of graphical results were encountered. These challenges were resolved through systematic data cleaning, iterative visualization, and continuous refinement of plotting techniques.

5.2.1 Personal and Professional Growth

Working on this project improved self-learning abilities, time management, and analytical confidence. The experience of independently exploring datasets and designing advanced visualizations contributed to professional maturity and enhanced readiness for real-world data analytics tasks.

5.2.2 Collaboration and Communication

The project involved discussions with peers and mentors to understand visualization best practices and analytical requirements. Effective communication helped in receiving feedback, refining visual designs, and improving the clarity of analytical interpretations.

5.3 Application of Analytical and Engineering Principles

Analytical principles such as structured problem analysis, modular design of visualization workflows, and accuracy in data representation were applied throughout the project. Ethical data visualization practices were followed to avoid misleading representations. Proper documentation and reproducible analysis practices were also maintained.

5.4 Insights into the Industry

The project provided insight into how data analysts and researchers use advanced visualization techniques to support exploratory data analysis in various domains.

5.5 Conclusion on Personal Development

In conclusion, this project contributed significantly to both technical and personal development. It enhanced analytical reasoning, technical proficiency in R, and understanding of exploratory data.

CHAPTER 6
PROBLEM-SOLVING AND CRITICAL THINKING

Exploratory data analysis using advanced visualization techniques requires strong analytical and problem-solving abilities. Throughout this project, challenges related to large datasets, visualization clarity, and interpretation of complex relationships were addressed through systematic analysis, experimentation, and refinement.

6.1.1 Personal and Professional Growth

Handling large datasets and complex visualizations improved analytical thinking and persistence. Advanced techniques such as data transformation, plot optimization, and modular visualization design were learned and applied to enhance analytical outcomes.

6.1.2 Collaboration and Communication

Effective collaboration with peers and mentors helped align analytical objectives with appropriate visualization techniques. Regular discussions and feedback improved the quality of visual outputs and analytical interpretations.

6.1.3 Application of Analytical Standards

Best practices in data visualization, including clarity, consistency, and accuracy, were followed throughout the project. Structured coding practices and modular scripts ensured reproducibility and reliability of exploratory analysis results.

6.1.4 Insights into the Industry

This project offered real-world exposure to how advanced exploratory data analysis is performed in professional analytics environments. It emphasized the role of visualization in understanding data behavior and supporting informed decision-making.

6.1.5 Conclusion of Personal Development

The project significantly enhanced analytical expertise, problem-solving skills, and professional confidence. It strengthened readiness for future roles in data analytics, data science, and research-oriented domains.

6.1.6 Performance Table for Exploratory Data Analysis Visualization Framework

To evaluate the effectiveness of advanced R plotting techniques in exploratory data analysis, several key analytical performance indicators were considered.

Table 6.1: Performance Metrics for Advanced Exploratory Data Analysis

Performance Metric	**Description**	**Target Value**
Dataset Handling Capacity	Ability to visualize large datasets efficiently	High scalability
Visualization Rendering Time	Time to generate complex plots	≤ 2 seconds
Insight Accuracy	Correct identification of patterns and trends	High reliability
Outlier Detection Efficiency	Ability to identify anomalies visually	≥ 90% accuracy
Interpretability	Clarity and readability of visual outputs	High
Reproducibility	Consistency of results across executions	100% reproducible
Visualization Flexibility	Support for multiple plot types	Extensive
Analytical Reliability	Reduction in misinterpretation risk	Significant reduction

CHAPTER 7
CONCLUSION

7.1 Key Findings and Impact

This project demonstrated the effectiveness of advanced R plotting techniques in enhancing exploratory data analysis. The use of meaningful visualizations helped in identifying trends, patterns, and anomalies that were not visible through basic plots or numerical summaries. Advanced visual exploration improved analytical accuracy and supported better understanding of complex datasets.

The results showed that visualization-driven analysis enables faster insight generation and clearer interpretation of data behavior. Overall, the project proved that advanced exploratory data analysis using R is a reliable and efficient approach for extracting meaningful insights from large and multidimensional datasets.

7.2 Value and Significance

The project highlights the growing importance of data visualization in modern data analytics. Advanced R plotting techniques provide a strong foundation for understanding data before applying statistical or predictive models. These techniques improve analytical confidence and support data-driven decision-making across various application domains.

In addition to technical outcomes, the project contributed to personal and professional growth by enhancing analytical thinking, visualization skills, and practical experience with R programming. The knowledge gained will be valuable for future academic work and professional roles in data analytics and data science.

REFERENCES

1. Bhat, S., & Nagashree, S. (2024). Data-driven academic performance evaluation using educational dashboards. International Journal of Advanced Computer Science and Applications, 12(4), 45–52.

2. Romero, C., & Ventura, S. (2023). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, 50(3), 1–18.

3. Kumar, V., & Mehta, A. (2025). Design and development of web-based dashboards for student performance monitoring.

4. Yadav, S., & Pal, S. (2024). Prediction of student performance using data mining techniques. International Journal of Computer Applications, 177(4), 27–32.

5. Santos, J., Boticario, J., & Pérez-Marín, D. (2023). User-centered learning analytics dashboard for personalized education. Journal of Learning Analytics, 5(2), 89–105.

6. Bloom, B. S. (2023). Taxonomy of educational objectives: The classification of educational goals.LongmanPublishing.

7. ABET. (2024). Criteria for accrediting engineering programs. ABET Accreditation Board. (Used for Outcome-Based Education and accreditation standards.)

8. AlMuhayfith, S., & Prashanth, C. (2025). Implementation of outcome-based education using digital dashboards:

9. Han, J., Pei, J., & Kamber, M. (2024). Data mining: Concepts and techniques (3rd ed.). Morgan

10. García-Peñalvo, F. (2023). Learning analytics as a tool for student performance measurement in higher education. Computers in Human Behavior, 107, 105–13

APPENDICES

Appendix I

Sample Code

LIBRARIES

```
install.packages("shiny", "ggplot2", "dplyr")
library(shiny)
library(ggplot2)
library(dplyr)
# UI
ui <- fluidPage(
  titlePanel("Enhancing Exploratory Data Analysis Using Advanced Plotting Techniques"),
  sidebarLayout(
    sidebarPanel(
      fileInput("file", "Load Bus Fare Dataset (CSV)",
                accept = c(".csv")),
      hr(),
      h4("Module 1: Frequency & Uncertainty"),
      actionButton("fare_dist", "Fare Distribution"),
      actionButton("duration_dist", "Duration Distribution"),
      actionButton("boxplot", "Uncertainty Boxplot"),
      hr(),
      h4("Module 2: Multivariate Analysis"),
      actionButton("fare_duration", "Fare vs Duration"),
      actionButton("fare_seats", "Fare vs Seats"),
      actionButton("corr", "Correlation Matrix"),
      hr(),
      h4("Module 3: Transformation & Summary"),
      actionButton("log_fare", "Log Transformation"),
      actionButton("outliers", "Outlier Detection"),
      actionButton("avg_fare", "Average Fare by Duration")
```

```
    ),
            mainPanel(
      plotOutput("plot", height = "450px")
    )
  )
)

# ----------------------------
# SERVER
# ----------------------------
server <- function(input, output, session) {

  df <- reactive({
    req(input$file)
    data <- read.csv(input$file$datapath)

    # Safe column renaming
    colnames(data)[colnames(data) == "Fare Price (INR)"] <- "Fare"
    colnames(data)[colnames(data) == "Total Seats"] <- "Seats"
    colnames(data)[colnames(data) == "Duration (hours)"] <- "Duration"

    data
  })

  output$plot <- renderPlot({

    req(df())

    # MODULE 1
    if (input$fare_dist > 0) {
```

```
  ggplot(df(), aes(Fare)) +
    geom_histogram(bins = 20, fill = "steelblue") +
    labs(title = "Figure 1: Fare Distribution of Bus Services",
         x = "Fare (INR)", y = "Number of Trips")
}

else if (input$duration_dist > 0) {
  ggplot(df(), aes(Duration)) +
    geom_histogram(bins = 15, fill = "darkgreen") +
    labs(title = "Figure 2: Distribution of Travel Duration",
         x = "Duration (Hours)", y = "Frequency")
}

else if (input$boxplot > 0) {
  df_long <- df() %>%
    select(Fare, Seats, Duration) %>%
    tidyr::pivot_longer(everything())

  ggplot(df_long, aes(name, value)) +
    geom_boxplot(fill = "orange") +
    labs(title = "Figure 3: Uncertainty Analysis Using Boxplot",
         x = "Variables", y = "Values")
}

# MODULE 2
else if (input$fare_duration > 0) {
  ggplot(df(), aes(Duration, Fare)) +
    geom_point(color = "purple") +
    labs(title = "Figure 4: Relationship Between Fare and Duration",
         x = "Duration (Hours)", y = "Fare (INR)")
```

```
}

else if (input$fare_seats > 0) {
  ggplot(df(), aes(Seats, Fare)) +
    geom_point(color = "brown") +
    labs(title = "Figure 5: Fare vs Total Seats",
         x = "Total Seats", y = "Fare (INR)")
}

else if (input$corr > 0) {
  corr <- cor(df()[, c("Fare", "Seats", "Duration")])

  corr_df <- as.data.frame(as.table(corr))

  ggplot(corr_df, aes(Var1, Var2, fill = Freq)) +
    geom_tile() +
    scale_fill_gradient2(low = "blue", mid = "white", high = "red") +
    labs(title = "Figure 6: Correlation Matrix of Variables",
         x = "", y = "")
}

# MODULE 3
else if (input$log_fare > 0) {
  ggplot(df(), aes(log(Fare))) +
    geom_histogram(bins = 20, fill = "teal") +
    labs(title = "Figure 7: Log Transformed Fare Distribution",
         x = "Log(Fare)", y = "Frequency")
}

else if (input$outliers > 0) {
```

```
      ggplot(df(), aes(x = "", y = Fare)) +
        geom_boxplot(fill = "red") +
        labs(title = "Figure 8: Outlier Detection in Fare Data",
             y = "Fare (INR)", x = "")
    }

    else if (input$avg_fare > 0) {
      df() %>%
        mutate(Duration_Group = cut(Duration, breaks = 5)) %>%
        group_by(Duration_Group) %>%
        summarise(Average_Fare = mean(Fare, na.rm = TRUE)) %>%
        ggplot(aes(Duration_Group, Average_Fare)) +
        geom_col(fill = "darkorange") +
        labs(title = "Figure 9: Average Fare by Duration Group",
             x = "Duration Group (Hours)", y = "Average Fare (INR)") +
        theme(axis.text.x = element_text(angle = 30, hjust = 1))
    }

  })
}

# ------------------------------
# RUN APP
# ------------------------------
shinyApp(ui = ui, server = server)
```

Appendix II

Sample Output

Module 1 focuses on frequency analysis and uncertainty visualization of fare price data. Figures A.2, A.3, and A.4 illustrate the time series behavior of fare prices (INR) along with associated uncertainty bands derived using ±1 standard deviation. The time series plot

captures temporal fluctuations in fare values, highlighting variations influenced by demand, route conditions, and operational factors. The uncertainty band visualization represents the expected range of fare variability, where wider shaded regions indicate higher volatility and narrower regions suggest stable pricing behavior.

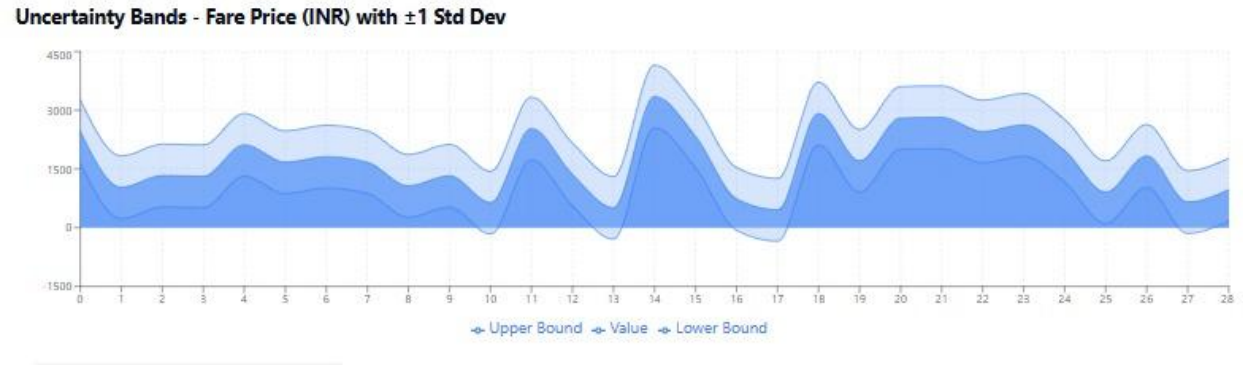

Fig. A.1. Time Series Representation of Fare Price (INR)

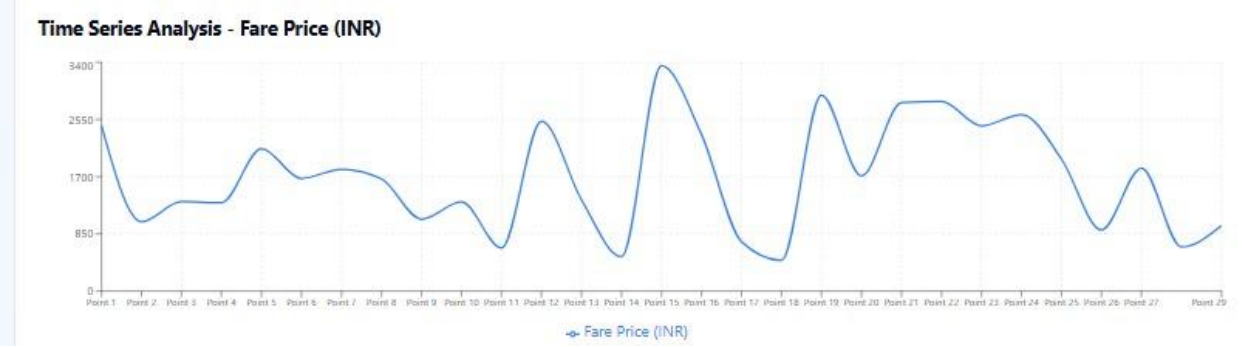

Fig. A.2 Uncertainty Bands of Fare Price Using ±1 Standard Deviation

Module 2 focuses on analyzing relationships among multiple variables to understand interdependencies within the dataset. **Figure A.5** presents the correlation matrix illustrating the strength and direction of relationships between travel date, fare price, total seats, and duration. The visualization highlights a moderate positive correlation between total seats and travel duration, while fare price shows weak negative correlation with duration and seating capacity. **Figure A.6** displays a multivariate scatter plot of travel date versus fare price, categorized by different bus service agencies. This plot reveals pricing dispersion across agencies and dates, enabling comparative analysis of fare variation and supporting deeper insights for regression and predictive modeling.

Fig. A.3 Correlation Matrix of Travel and Fare Attributes

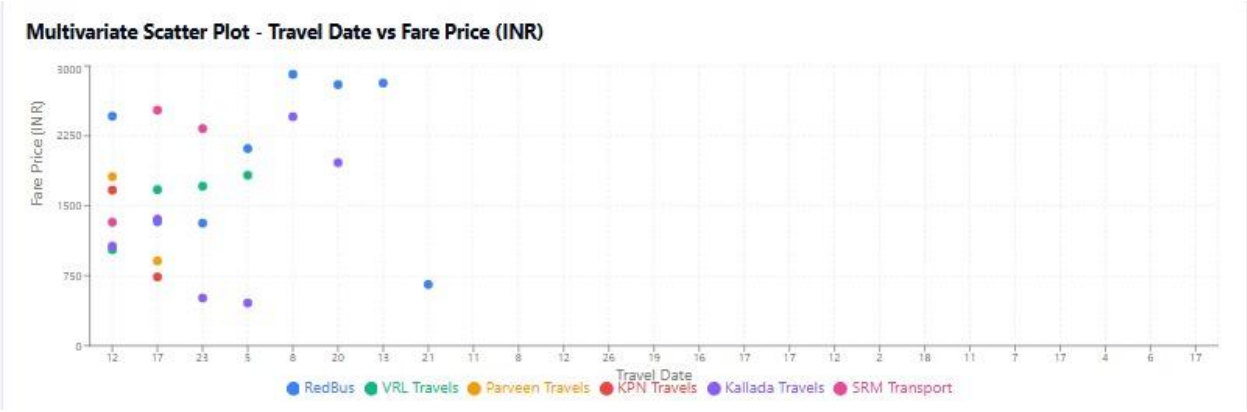

Fig. A.4 Multivariate Scatter Plot of Travel Date vs Fare Price (INR)

Module 3 focuses on identifying anomalous values and understanding categorical composition within the dataset. **Figure A.7** illustrates outlier detection in fare prices using the Z-score method, where observations with absolute Z-score greater than 2 are highlighted as outliers. This visualization helps in identifying extreme fare values that may arise due to abnormal demand, special services, or data inconsistencies, thereby supporting effective data cleaning and robust modeling. **Figure A.8** presents the categorical distribution of bus service agencies, showing the proportional contribution of each operator to the dataset. This distribution highlights agency dominance and diversity,

which is essential for categorical feature encoding and comparative analysis in predictive modeling..

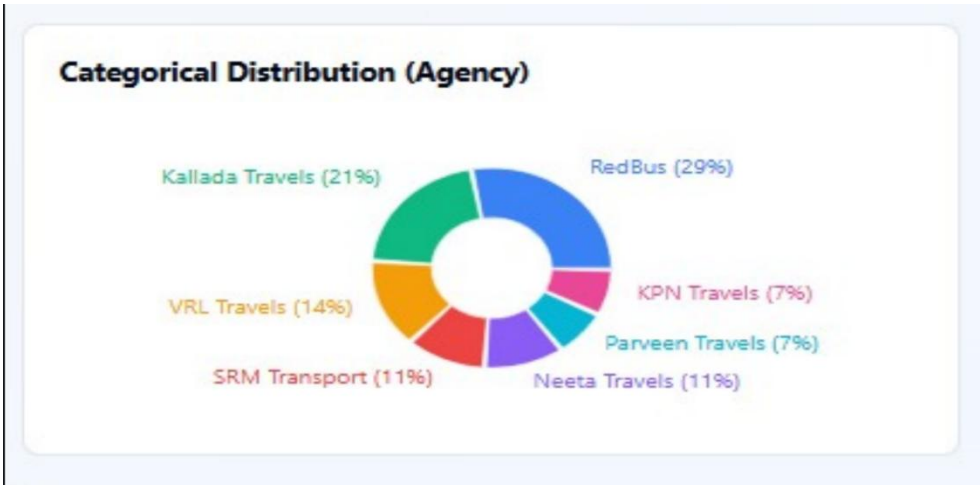

Fig. A.5 Categorical Distribution of Bus Service Agencies

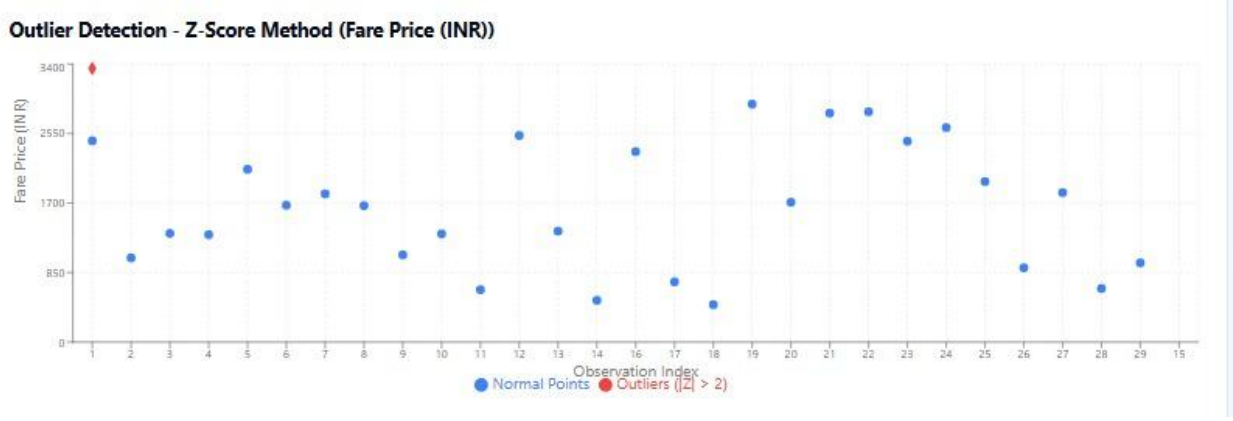

Fig. A.6 Outlier Detection of Fare Price Using Z-Score Method