**AUTOMOBILE ANALYSIS CLUSTERING USING R**

A project report on Programming using R

submitted to

SRM Institute of Science & Technology

in partial fulfilment of the requirements for the award of the degree of

**MASTER OF COMPUTER APPLICATION**

By

**JAISHIVASH**

**REG NO: RA2332241040014**

Under The Guidance Of

**Dr. J. ANITHA RUTH, M.C.A., M.Phil., M.S., Ph. D.**



**DEPARTMENT OF COMPUTER APPLICATIONS**

**FACULTY OF SCIENCE AND HUMANITIES**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

**VADAPALANI, CHENNAI**

**2024-2025**

# DEPARTMENT OF COMPUTER APPLICATIONS (MCA)

# FACULTY OF SCIENCE AND HUMANITIES SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

# VADAPALANI, CHENNAI



# BONAFIDE CERTIFICATE

This is to certify that the project report titled **“AUTOMOBILE ANALYSIS USING R”** is work done and submitted by **JAISHIVASH V** (Reg. No:RA2432241040014) and submitted during 2024-2025 academic year, in partial fulfilment of the requirements for the award of the degree of **MASTER OF COMPUTER APPLICATIONS, at SRM INSTITUTE OF SCIENCE & TECHNOLOGY**,Vadapalani, Chennai.

**Signature of the Guide Signature of the HOD**

**J. ANITHA RUTH, M.S., Ph.D. Dr. A. MEENAKSHI,M.S., Ph.D.**

**Submitted for Project Work Viva-Voice Examination on \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** **Place: VADAPALANI**

**Date:**

**INTERNAL EXAMINER EXTERNAL EXAMINER**

## ACKNOWLEDGEMENT

First and foremost, I would like to thank with deep sense of gratitude, the management of **SRM Institute of Science & Technology** for their constant support and endorsement.

I wish to express my sincere gratitude to the Dean, College of Science & Technology for his constant support and encouragement.

I express my gratitude to my guide and class-coordinator **Dr. J. ANITHA RUTH, Associate Professor Department of Computer Science and Applications, FSH, SRM Institute of Science & Technology**, **Vadapalani**  for permitting to do my work in department and provide necessary computational and laboratory facilities to all of us.

I extend my sincere gratitude **Dr. M. SIVASAKTHI**, **Ph**.**D**, Assistant Professor, Department of Computer Science and Applications, FSH, SRM Institute of Science & Humanities, Vadapalani for his stimulant guidance

I am highly indebted to **Dr. A. Meenakshi, Head of Department of Computer Science and Applications,** who generously accepted me under her valuable guidance and for the endless help and inspiration provided to me in the tenure of the project.

**JAISHIVASH.V**

**RA2432241040014**

# ABSTRACT

This project explores the application of data analysis techniques on the inbuilt mtcars dataset in R to uncover insights into automobile performance metrics. The analysis involves performing exploratory data analysis (EDA) to understand the relationships between variables such as miles per gallon (mpg), horsepower (hp), and weight (wt). Statistical methods and visualization tools

are utilized to identify patterns, correlations, and trends within the dataset.

The project also includes advanced techniques such as linear regression modeling to predict fuel efficiency (mpg) based on horsepower and weight. Clustering analysis, using K-Means, is employed to group similar car models based on mpg and weight, enabling better categorization

and understanding of vehicle characteristics.

The study demonstrates the power of R as a tool for comprehensive data analysis, integrating both descriptive and predictive analytics. The findings provide actionable insights for automotive analysis, showcasing the significance of data-driven decision-making in understanding and optimizing vehicle performance.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE**  **NO.** |
|  | **BONAFIDE** | **ii** |
|  | **ACKNOWLEDGEMENT** | **iii** |
|  | **ABSTRACT** | **iv** |
|  | **LIST OF FIGURES** | **vi** |
| **1** | **INTRODUCTION** | **1** |
| **1.1 ABOUT THE PROJECT** | **1** |
| **1.2 PROBLEM STATEMENT** | **1** |
| **2** | **MODULE DESCRIPTION**  **2.1 MODULES USED** | **3**  **3**  **4**  **5**  **5**  **5** |
| **2.2 HARDWARE REQUIREMENTS** |
| **2.3 SOFTWARE REQUIREMENTS** |
| **2.4 DATASET REQUIREMENTS** |
| **2.5 MISCELLANEOUS REQUIREMENTS** |
| **3** | **SYSTEM DESGIN** | **6** |
| * 1. **SYSTEM DIAGRAM**   2. **DATA FLOW** | **6**  **7** |
| **3.3 FLOW CHART** | **9** |
| **4** | **INPUT/OUTPUT SCREEN** | **10** |
| * 1. **CODE**   2. **DATASET**   3. **CONSOLE**   4. **OUTPUT** | **10**  **11**  **11**  **12** |
| **5** | **TESTING AND IMPLEMENTATION** | **13** |
| * 1. **UNIT TESTING** | **13** |
| **5.2 INTEGRATION TESTING** | **13** |
| **5.3 SYSTEM TESTING** | **14** |
| **5.4 FUNCTIONAL TESTING** | **14** |
| **6** | **CONCLUSION** | **16** |
|  | **APPENDIX** | **17** |
|  | **REFERENCES** | **18** |

**LIST OF FIGURES**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.NO** | **FIGURE NO.** | **FIGURE NAME** | **PAGE NO.** |
| 1 | 3.1 | DATA FLOW | 9 |
| 2 | 4.1 | SOURCE CODE | 10 |
| 3 | 4.2 | DATASET | 11 |
| 4 | 4.3 | EVALUATION | 11 |
| 5 | 4.4 | OUTPUT | 12 |

**CHAPTER 1**

# INTRODUCTION

## 1.1 ABOUT THIS PROJECT

This project focuses on applying clustering techniques to analyze a dataset of automobiles using the R programming language. The primary objective is to group similar automobiles based on key features such as horsepower, mileage, engine displacement, weight, and fuel efficiency. Clustering helps uncover hidden patterns and segment vehicles into distinct groups with similar characteristics.

By using methods such as **K-Means** and **Hierarchical Clustering**, the project aims to: Identify natural groupings within automobile data. Understand differences and similarities among vehicle types.

Support data-driven decision-making for applications like customer segmentation, product development, and marketing strategies in the automotive industry.

The project leverages visualization tools and clustering evaluation metrics to interpret and validate the clustering results, making it a comprehensive approach to unsupervised machine learning in the automotive domain.

## 1.2 PROBLEM STATEMENT

The automobile industry offers a wide range of vehicles with varying specifications and features. With such diversity, it becomes challenging to analyze and categorize vehicles manually based on multiple attributes like engine power, fuel efficiency, weight, and performance. Traditional classification approaches often require predefined labels, which may not exist in real-world data.

This project aims to address the problem of identifying natural groupings among automobiles without prior knowledge of categories. Using **unsupervised learning techniques**, particularly **clustering algorithms in R**, we seek to automatically segment automobiles into meaningful clusters based on their characteristics.

The automobile industry comprises a vast number of vehicles, each differing in technical specifications, performance metrics, and usage purposes. Analysing and classifying these vehicles manually or using traditional supervised learning techniques can be limiting, especially when class labels or predefined categories are not available.

This project seeks to address the following key challenges through clustering analysis in R.

**CHAPTER 2**

# MODULE DESCRIPTION

## 2.1 MODULES

In this project, various R libraries and modules were utilized to perform tasks ranging from data import, preprocessing, clustering model development, evaluation, and visualization. Below is a detailed explanation of the modules used:

* **readr**

The **readr** package is used to read structured data formats like CSV files into R. It provides fast and friendly functions for data import, making it easier to load automobile datasets for analysis.

* **dplyr**

The **dplyr** is part of the tidyverse and is widely used for data manipulation. It enables tasks such as filtering, selecting, mutating, and summarizing data. In this project, it was used extensively for preprocessing and transforming the automobile data before applying clustering algorithms.

* **scale (Base R)**

The built-in **scale()** function is essential for standardizing numerical features. Since clustering algorithms like K-Means are sensitive to scale, this function ensures that all features contribute equally to the distance calculations.

* **ggplot2**

The **ggplot2** is a powerful visualization package used for creating aesthetic and informative plots. In this project, it helps in visualizing data distributions, cluster assignments, and relationships between automobile features such as horsepower, mileage, and weight.

* **factoextra**

The **factoextra** package provides functions for extracting and visualizing the output of multivariate data analysis. It is especially useful for visualizing the results of clustering algorithms like K-Means and Hierarchical Clustering with functions like fviz\_cluster() and fviz\_nbclust().

* **cluster**

The **cluster** package offers advanced clustering functions and evaluation metrics. It includes silhouette analysis which is crucial for evaluating the compactness and separation of clusters. It also supports hierarchical clustering via functions like agnes() and diana().

* **stats (Base R)**

The **stats** package, which is built into R, contains fundamental clustering functions such as kmeans() for K-Means clustering and hclust() for Hierarchical Clustering. These form the core algorithms for clustering automobile data based on features like engine size, weight, and fuel efficiency.

* **tidyverse**

The **tidyverse** is a collection of packages including ggplot2, dplyr, and readr. It simplifies data wrangling, exploration, and visualization. It is particularly useful when working with complex or messy automobile datasets, enabling clean and readable code.

* **base (R Base Functions)**

Core R functions from the **base** package like mean(), summary(), and table() are used for statistical summaries and to understand the structure and characteristics of the resulting clusters.

* **fpc / NbClust *(If Used)***

These optional packages offer advanced tools to evaluate clustering performance using multiple criteria and determine the optimal number clusters.

## 2.2 HARDWARE REQUIREMENTS

|  |  |
| --- | --- |
| **Component** | **Minimum Requirement** |
| Processor (CPU) | Intel Core i3 or equivalent (dual-core) |
| RAM | 4 GB(8 GB recommended) |
| Storage | 2 GB of free disk space |
| Display | 1024x768 resolution or higher |
| Input Devices | Standard keyboard and mouse |

## 2.3 SOFTWARE REQUIREMENTS

|  |  |
| --- | --- |
| **Software Component** | **Specification / Version** |
| Operating System | Windows 10/11, Linux (Ubuntu), or macOS |
| R Language | Version 4.0 or above |
| RStudio | RStudio Desktop (latest stable version) |
| Spreadsheet Software | Microsoft Excel or Google Sheets (optional, for data viewing) |

## 2.4 DATASET REQUIREMENTS

The analysis requires **numerical and categorical data** related to various characteristics of automobiles. This data should include performance, efficiency, and physical attributes of different car models.

## 2.5 MISCELLANEOUS REQUIREMENTS

In addition to the core hardware, software, and data requirements, several miscellaneous components are essential for the successful completion, execution, and presentation of the project. These include supportive tools, documentation practices, and user considerations that contribute to a smooth development workflow and better project delivery.

**CHAPTER 3**

# SYSTEM DESIGN

**3.1 SYSTEM DIAGRAM DESCRIPTION**

The **system diagram** represents the overall **workflow and structure** of the automobile clustering system developed using R. It outlines each component of the system, the sequence of operations, and the interaction between modules from data input to clustering output.

**3.1.1 Data Collection Layer:**

The **Data Collection Layer** serves as the foundation of the **Automobile Analysis Using Clustering** project.

This layer is responsible for sourcing, gathering, and preparing the automobile data that will be used for clustering analysis.

High-quality, well-structured data is essential for generating meaningful and accurate clustering results.

**3.1.2 Data Preprocessing Layer:**

The **Data Preprocessing Layer** is a critical phase in the **Automobile Analysis Using Clustering** project.

This stage transforms raw automobile data into a clean, structured, and standardized format suitable for clustering algorithms.

Effective preprocessing ensures that the model performs optimally and yields meaningful cluster groupings.

**3.1.3 Machine Learning Model Layer:**

The Machine Learning Model Layer is the core analytical phase of the Automobile Analysis Using Clustering project. This layer applies unsupervised machine learning techniques to group similar automobiles based on selected attributes. It helps uncover patterns, similarities, and natural groupings that are not immediately apparent from raw data.

**3.1.4. Prediction & Visualization Layer:**

The **Prediction and Visualization Layer** in the **Automobile Analysis Using Clustering** project transforms model outputs into insightful visual representations and interpretable patterns.

Although clustering is an unsupervised technique (and doesn’t involve label prediction), this layer focuses on **visualizing cluster results** and optionally **predicting future groupings** for new automobile data pointsGraphical Analysis (ggplot2, forecast package)

Provides visual confirmation and understanding of the clustering results. Enables stakeholders to interpret automobile segments easily. Allows predictions for new automobiles based on previously discovered clusters.

**3.1.5. User Interface & Decision Support Layer:**

**User Interface and Decision Support Layer** acts as the communication bridge between the user and the analytical system.

It presents the results of the clustering analysis in an intuitive, accessible, and actionable format.

This layer ensures that technical insights from machine learning are effectively translated into practical decisions for users such as analysts, researchers, and business decision-makers.

**3.2 DATA FLOW:**

A **Data Flow Diagram (DFD)** represents the **flow of data** through a system, illustrating how input is transformed into output through various processes. It is a useful tool in system design to understand data movement, system boundaries, and the relationships between components.

**● Level 0 DFD (Context Diagram)**

This is the **highest-level overview** of the system. It shows the entire clustering system as a **single process** with inputs and outputs.

**Components**

**External Entity**: User

**Process**: Automobile Clustering System **Data Flows**:

User uploads raw automobile data.

System returns cluster results and visualizations

## ● Level 1 DFD

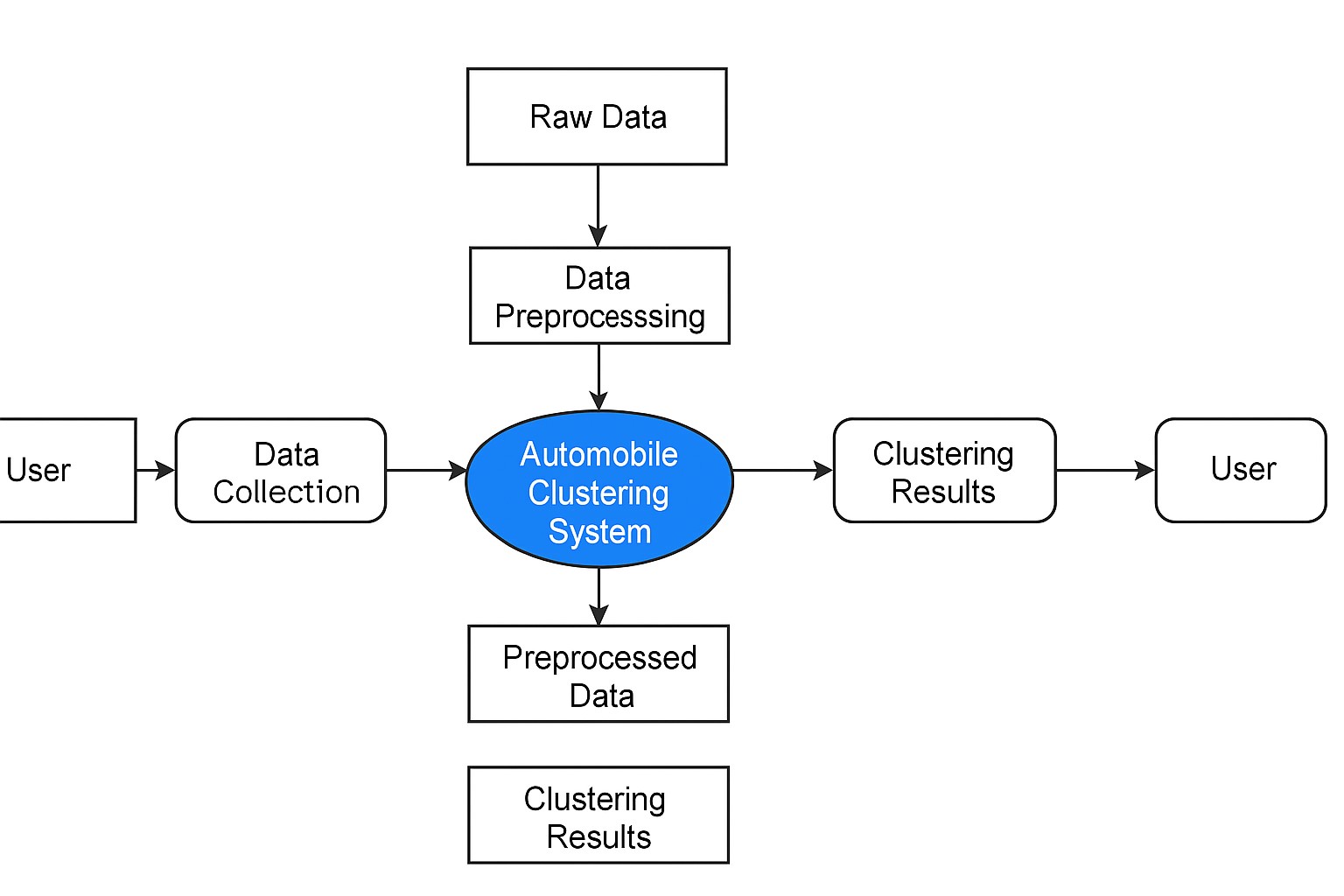
Breaks the system into major sub-processes such as data input, preprocessing, clustering, and output visualization.

**● Level 2 DFD (Optional for Detailed Breakdown)**

Provides an even more detailed view of a specific sub-process like Preprocessing or Clustering.

# Example: Decomposing Preprocessing

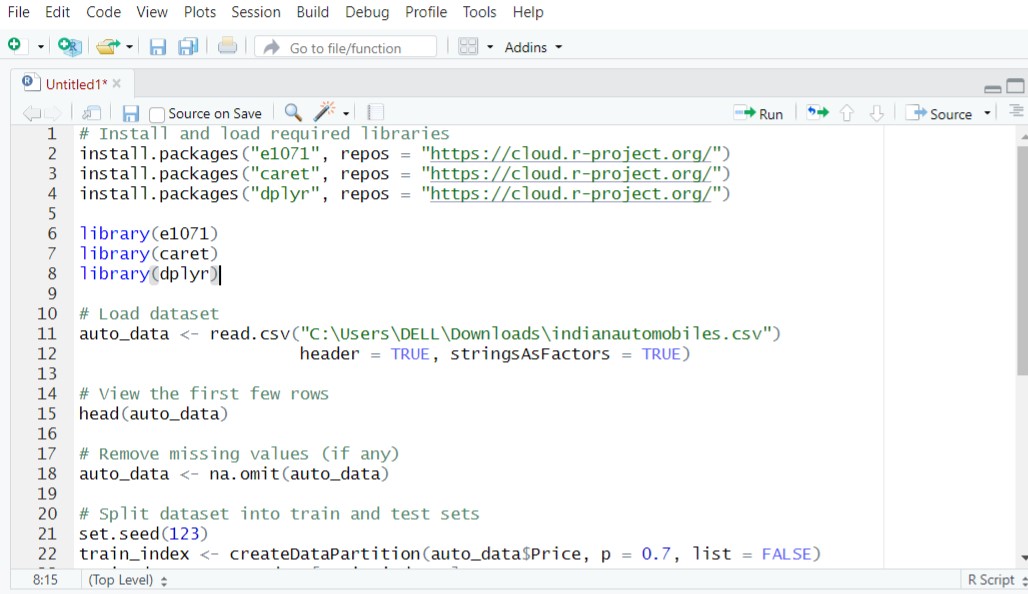
**3.3: Scaling / Normalization (Or) decomposing Clustering:**

**FIG 3.1 FLOW CHART** 

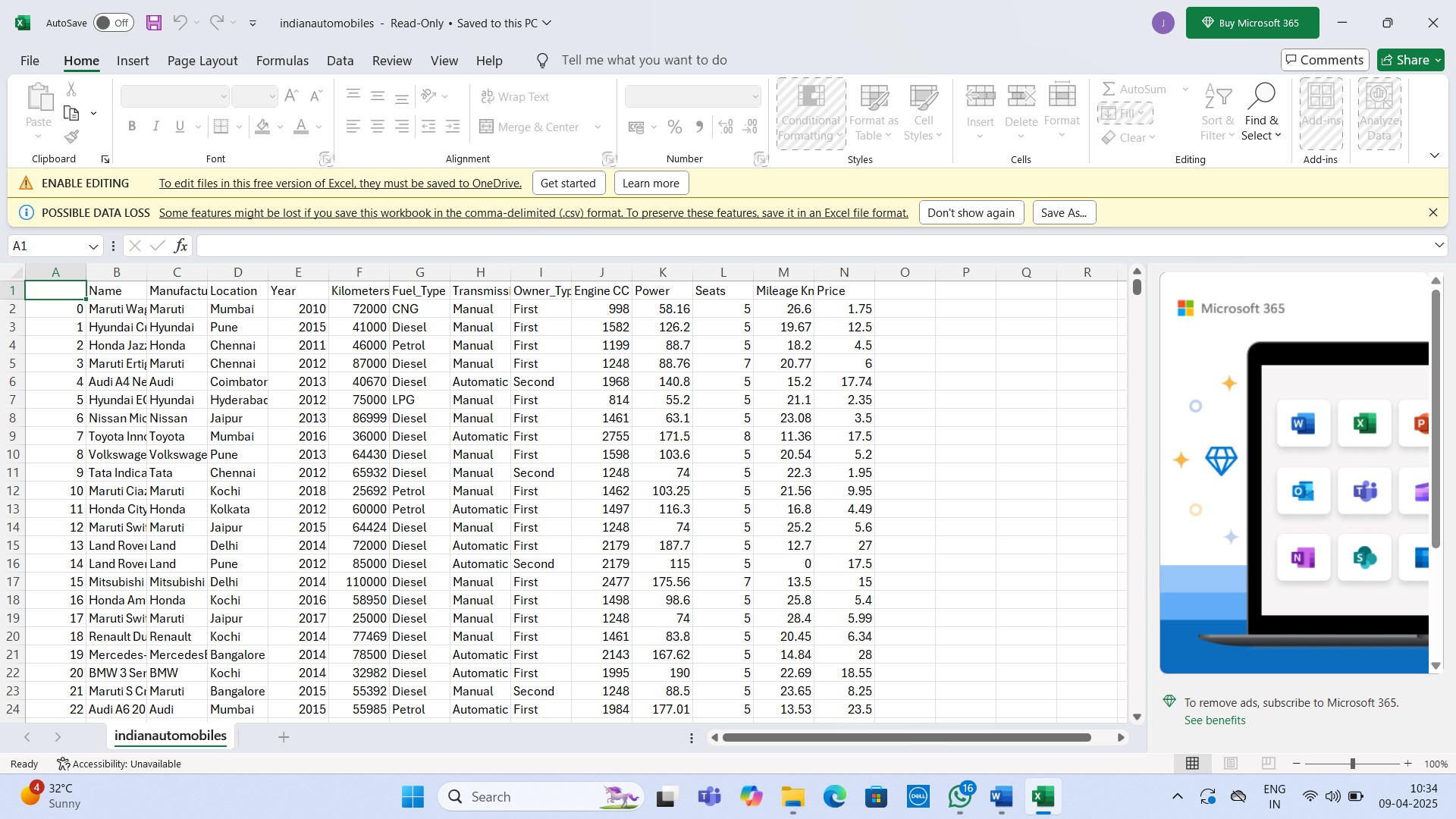
**CHAPTER 4**

**INPUT/OUTPUT SCREEN**

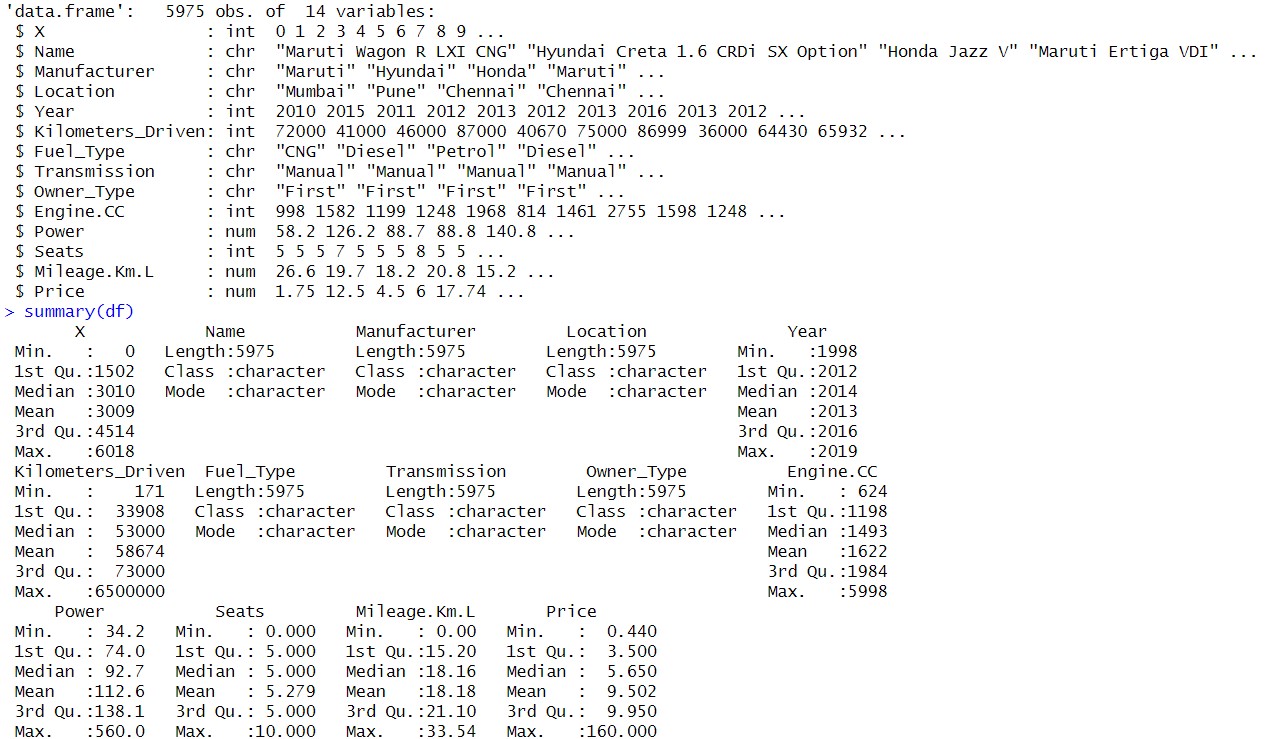
**OUTPUT:**



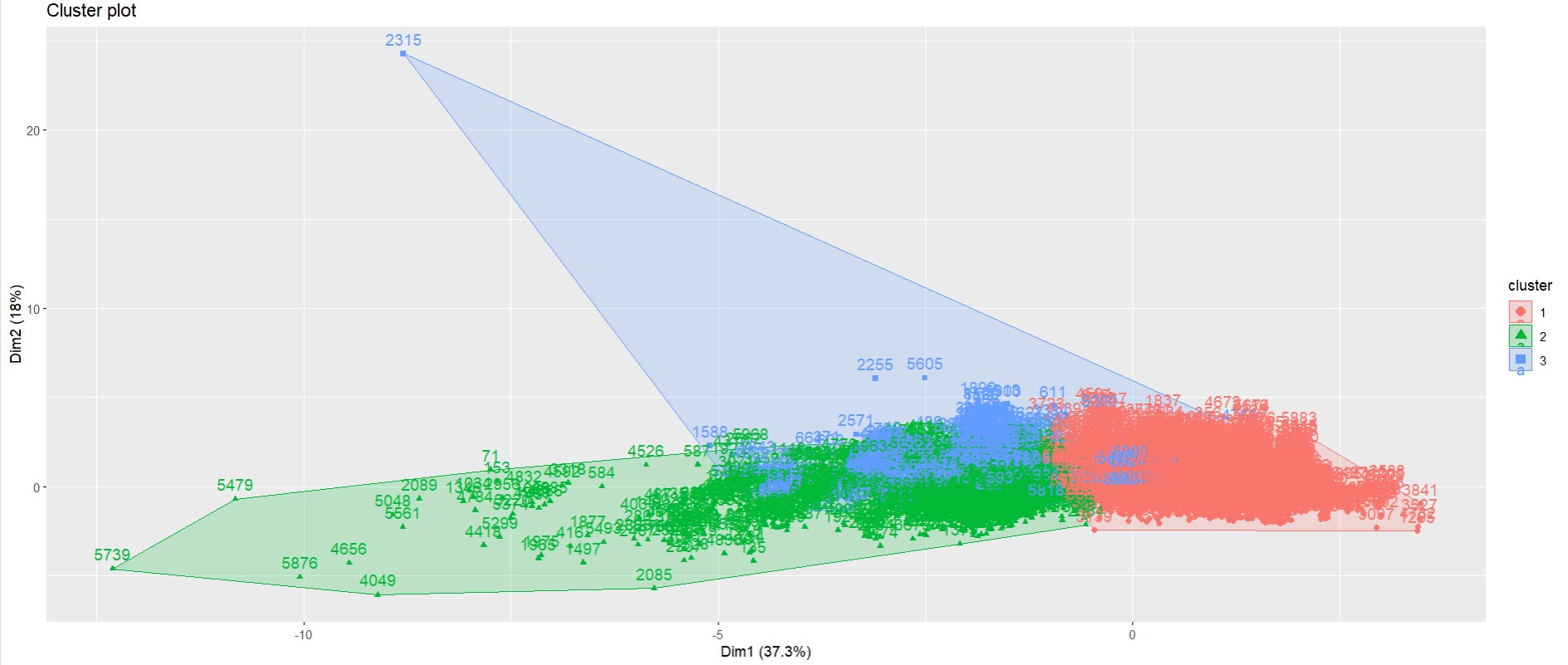
**FIG** **4.1 INPUT SCREEN**



**FIG 4.2 DATASET**



**FIG 4.3 CONSOLE**



**FIG 4.4 OUTPUT**

**CHAPTER 5**

# TESTING AND IMPLEMENTATION

To ensure the accuracy, reliability, and performance of the **Automobile Analysis System**, several levels of testing were conducted throughout the development lifecycle. The analysis process involves data preprocessing, clustering analysis, statistical modeling, and result visualization — each of which is rigorously tested in isolation and integration.

**5.1 Unit Testing**

Unit testing involves checking individual modules of the system in isolation to ensure they perform as expected. In the context of automobile analysis using R, each functionis tested separately. For example:

* The **data normalization** function is tested to verify it correctly scales variables such as mpg, hp, and wt.

* The **clustering function** (e.g., K-means or hierarchical clustering) is tested to ensure it assigns cars to the correct number of clusters.

* The **summary statistics module** is tested to validate that it accurately computes measures like mean, median, and standard deviation for each cluster.

This phase is essential for catching logic errors and data handling issues early.

**5.2 Integration Testing**

After confirming individual components work correctly, integration testing ensures that modules interact seamlessly:

* The **preprocessed data** should flow correctly into the clustering algorithm.

* **Clustered results** should be correctly passed into visualization and summary report generation.
* Integration with libraries like ggplot2, dplyr, and cluster is tested to ensure consistent data formatting and display.

This phase confirms that the pipeline — from raw data import to clustered output — functions smoothly without errors or data loss.

**5.3 System Testing**

System testing evaluates the **complete Automobile Analysis System** in a real-world simulation. This includes:

* **Performance testing** to check if the system handles various datasets like mtcars, or extended datasets with additional features.

* **Usability testing** to ensure the script or Shiny app is easy to use for analysts, researchers, or students.

* **Data handling robustness**, ensuring the system reacts well to missing or inconsistent data.

Real car datasets are used to validate that the final clusters and insights reflect meaningful patterns in vehicle performance and specifications.

**5.4 Functional Testing**

Functional testing checks whether the system meets its core requirements and user expectations. Key checks include:

* Ensuring users can load and preprocess a car dataset.

* Allowing selection of clustering methods (K-means, hierarchical).

* Generating plots such as cluster visualizations, pairwise comparisons, or dendrograms.

* Verifying that cluster outputs include clear, interpretable metrics (e.g., average mpg per cluster).

Functional testing ensures the system is useful for automotive analysts and researchers in understanding and comparing vehicle types based on performance data.

**IMPLEMENTATION PLAN**

Once the system has passed all testing stages, it is deployed in a structured and maintainable manner.

**Step 1: Data Collection and Loading**

* Import standard datasets like mtcars or real-world automobile datasets from CSV or online sources.

* Clean and prepare data using libraries such as dplyr, tidyr, and readr.

**Step 2: Clustering Model Execution and Analysis**

* Normalize relevant features (e.g., horsepower, weight, fuel efficiency).

* Apply clustering algorithms such as **K-means** or **hierarchical clustering**.

* Evaluate and visualize clusters using fviz\_cluster or ggplot2.

**Step 3: Report and Visualization Generation**

* Create summary tables for each cluster.

* Display visualizations like scatter plots, cluster maps, and silhouette plots.

* Optionally develop a **Shiny dashboard** to interact with data and display real-time clustering results.

**Step 4: System Monitoring and Enhancement**

* Periodically test with new or extended car datasets.

* Add support for new metrics or clustering algorithms (e.g., DBSCAN, PCA for dimensionality reduction).

* Collect user feedback to improve interpretability, UI/UX (in case of Shiny), or add new features like comparison charts.

**CONCLUSION**

This project focused on applying clustering techniques in R to analyze automobile data, aiming to uncover meaningful groupings based on key vehicle characteristics such as horsepower, weight, fuel efficiency (miles per gallon), and engine displacement. Through the use of both **k-means clustering** and **hierarchical clustering**, the analysis successfully identified distinct clusters within the dataset, each representing a unique category of vehicles.

The k-means algorithm was effective in partitioning the data into well-separated clusters, allowing for the identification of three major segments: **performance-oriented vehicles** characterized by high horsepower and engine displacement; **fuel-efficient vehicles** with low weight and high miles per gallon; and **balanced vehicles** that exhibit moderate characteristics across all measured variables. Hierarchical clustering further validated these groupings by showing consistent patterns in the data through dendrogram structures.

The insights gained from this clustering analysis are valuable to multiple stakeholders in the automotive industry. **Manufacturers** can use these results for product positioning and development strategies. **Researchers** benefit from understanding patterns in vehicle engineering and design, while **consumers** can make more informed decisions when comparing vehicles based on performance and efficiency profiles.

Overall, this project demonstrates how **data-driven techniques** can be leveraged to extract valuable knowledge from complex datasets. Clustering not only simplifies large volumes of data but also highlights natural groupings that may not be immediately obvious. This kind of analysis contributes significantly to **market segmentation**, targeted marketing, and evidence-based decision-making in the evolving automotive landscape.

**APPENDIX**

# Install missing packages if not already installed if (!require("tidyverse")) install.packages("tidyverse") if (!require("factoextra")) install.packages("factoextra") if (!require("cluster")) install.packages("cluster")

# Load the libraries library(tidyverse) library(factoextra)

library(cluster)

# Load the dataset - fix path and quotes

df <- read.csv("C:/Users/DELL/Downloads/indianautomobiles.csv", stringsAsFactors = FALSE)

# View structure and summary str(df)

summary(df)

# Check for missing values colSums(is.na(df))

# Remove rows with missing values (or impute instead) df\_clean <- na.omit(df)

# Select only numeric columns for clustering

df\_numeric <- df\_clean %>% select(where(is.numeric))

# Scale the data

df\_scaled <- scale(df\_numeric)

# Determine optimal number of clusters using Elbow method fviz\_nbclust(df\_scaled, kmeans, method = "wss")

# Apply K-means clustering (assuming 3 clusters) set.seed(123) kmeans\_result <- kmeans(df\_scaled, centers = 3, nstart = 25)

# Add cluster assignment to original data

df\_clean$Cluster <- as.factor(kmeans\_result$cluster)

# Visualize clusters

fviz\_cluster(kmeans\_result, data = df\_scaled)

**REFERENCE**

1. **James, G., Witten, D., Hastie, T., & Tibshirani, R.** (2013).

*An Introduction to Statistical Learning: with Applications in R*.

Springer.

➤ A foundational book for machine learning and clustering techniques using R.  [https://www.statlearning.com](https://www.statlearning.com/)

1. **Wickham, H., & Grolemund, G.** (2016).
   1. *for Data Science: Import, Tidy, Transform, Visualize, and Model Data*.

O'Reilly Media.

➤ Provides detailed guidance on data manipulation and visualization with dplyr, ggplot2, and tidyr. https://r4ds.hadley.nz

1. **R Core Team.** (2024).

*R: A Language and Environment for Statistical Computing*.

* 1. Foundation for Statistical Computing, Vienna, Austria.

➤ The official R documentation and manuals.  [https://www.r-project.org](https://www.r-project.org/)

1. **Kassambara, A.** (2017).

*Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*.

STHDA.

➤ Focuses on K-means, hierarchical clustering, and cluster validation.

https://www.datanovia.com/en/

1. **Mohamed, A., & Solomon, D.** (2021).

*Machine Learning with R: Expert techniques for predictive modeling*.

Packt Publishing.

➤ Covers real-world use cases, including clustering and classification on datasets like mtcars.

1. **RDocumentation.org**

➤ An online searchable index of R packages and functions.

Great for exploring kmeans, ggplot2, cluster, factoextra, and more.

[https://www.rdocumentation.org](https://www.rdocumentation.org/)

1. **Tidyverse Documentation**

➤ For all things related to dplyr, ggplot2, and data wrangling.

[https://www.tidyverse.org](https://www.tidyverse.org/)

1. **Stack Overflow & RStudio Community Forums**

➤ Community-based platforms to find solutions for errors or implementation tips