# ANALYSING EXAM SCORES USING DESCRIPTIVE STATISTICS

A PROJECT REPORT

*Submitted by*

**PADMAPRIYA S (2303811724322080)**

*in partial fulfillment of requirements for the award of the course*

**AGB1252 - 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 **“ANALYSING EXAM SCORES USING DESCRIPTIVE ANALYSIS”** is the bonafide work of **PADMAPRIYA S (2303811724322080)** who carried out the project work during the academic year 2024 - 2025 under my supervision.

| | |
|---|---|
| **SIGNATURE** | **SIGNATURE** |
| Dr.T. AVUDAIAPPAN, M.E.,Ph.D., | Ms.S.Murugavalli., M.E.,(Ph.D)., |
| **HEAD OF THE DEPARTMENT** | **SUPERVISOR** |
| PROFESSOR | ASSISTANT PROFESSOR |
| Department of Artificial Intelligence | Department of Artificial Intelligence |
| K.Ramakrishnan College of Technology (Autonomous) | K.Ramakrishnan College of Technology (Autonomous) |
| Samayapuram–621112. | 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 **“ANALYSING EXAM SCORES USING DESCRIPTIVE ANALYSIS”** is the result of original work done by me and best of our 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 **AGB1252 - FUNDAMENTALS OF DATA SCIENCE USING R**

.

**Signature**

PADMAPRIYA S

Place: Samayapuram

Date: 02.06.2025

# 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 my study in college.

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

Student academic performance is analyzed using descriptive statistics, clustering, regression, and classification techniques on a dataset containing marks from multiple subjects. Initial analysis explores score distributions and correlations between subjects. K-Means clustering identifies distinct groups of students with similar performance patterns. A linear regression model predicts total marks based on individual subject scores, revealing the impact of each subject. Additionally, a logistic regression classifier predicts pass/fail results, enabling early identification of students at risk. The combined methods provide valuable insights for educators to better understand performance trends&interventions.

## ABSTRACT WITH POs AND PSOs MAPPING

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

| ABSTRACT | POs MAPPED | PSOs MAPPED |
|---|---|---|
| Student academic performance is analyzed using descriptive statistics, clustering, regression, and classification techniques on a dataset containing marks from multiple subjects. Initial analysis explores score distributions and correlations between subjects.<br>K-Means clustering identifies distinct groups of students with similar performance patterns | **PO1 -3**<br>**PO2 -2**<br>**PO3 -3**<br>**PO4 -1**<br>**PO5 -3**<br>**PO6 -1**<br>**PO7 -3**<br>**PO8 -3**<br>**PO9 -2**<br>**PO10 -3**<br>**PO11-3** | **PSO1 -3**<br>**PSO2 -2** |

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

## TABLE OF CONTENTS

| CHAPTER NO. | TITLE | PAGE NO. |
|---|---|---|
| | **ABSTRACT** | **viii** |
| **1** | **INTRODUCTION** | **1** |
| | 1.1 Objective | 1 |
| | 1.2 Overview | 1 |
| | 1.3 Data Science related concepts | 2 |
| **2** | **PROJECT METHODOLOGY** | **3** |
| | 2.1 Proposed Work | 3 |
| | 2.2 Block Diagram | 5 |
| **3** | **MODULE DESCRIPTION** | **6** |
| | .3.1 Data Input Module | 6 |
| | 3.2 Data Preprocessing Module | 6 |
| | 3.3 Statistical Computational Module | 6 |
| | 3.4 Data Visualization Module | 6 |
| | 3.5 ML Algorithm Module | 6 |
| **4** | **CONCLUSION & FUTURE SCOPE** | **7** |
| | **APPENDIX A SOURCE CODE** | **8** |
| | **APPENDIX B SCREENSHOTS** | **10** |
| | **REFERENCES** | **12** |

# CHAPTER 1

# INTRODUCTION

## 1.1 INTRODUCTION

Analyzing exam scores is essential for understanding student performance and enhancing educational outcomes. This study leverages R programming to apply descriptive statistics, K-Means clustering, regression modeling, and classification techniques on student exam data. Descriptive analysis summarizes key performance metrics, while clustering groups students based on score patterns. Regression identifies influential factors and predicts scores, and classification categorizes students into performance levels. Together, these methods provide a comprehensive approach to uncovering insights and supporting data-driven academic strategies.

## 1.2 OBJECTIVE

The primary objective of this project is to analyze student exam scores using a combination of statistical and machine learning techniques in R. The specific goals include:

- To summarize and interpret exam data using descriptive statistical methods.
- To group students based on performance patterns using K-Means clustering.
- To identify key factors influencing exam scores through regression analysis.
- To classify students into performance categories (e.g., high, average, low) using classification models.
- To provide actionable insights that can support academic planning and targeted interventions.

# 1.3 DATA SCIENCE RELATED CONCEPTS

## 3.1 Data Import and Export

Functions like read.csv(), write.csv(), read_excel() to load and save data files.

### 3.1.1 Data Cleaning and Preprocessing

- Handling missing values using na.omit(), is.na(), replace().
- Data type conversion with as.numeric(), as.factor().
- Feature scaling using scale() and normalization.

## 3.2 Descriptive Statistics

- Summary functions: summary(), mean(), median(), sd(), var(), quantile().
- Data exploration: table(), cor(), aggregate().

## 3.3 Data Visualization

- Base R plotting: plot(), hist(), boxplot(), barplot().
- ggplot2 package: ggplot(), geom_point(), geom_bar(), geom_boxplot().

### 3.3.1 K-Means Clustering

- Function: kmeans() for clustering analysis.
- Finding optimal clusters using the Elbow Method or Silhouette Score.
- Visualization with fviz_cluster() from the factoextra package.

### 3.3.2 Regression Modeling

- Linear regression using lm() function.
- Model evaluation: summary(), residual plots, $R^2$, and RMSE.
- Diagnostic checks for multicollinearity using vif() from car package.

### 3.3.3 Classification Models

- Decision Tree using rpart() and rpart.plot().
- Logistic Regression using glm(family = "binomial").
- Splitting data: sample(), createDataPartition() from caret.
- Model evaluation: confusion matrix with table() or confusionMatrix().

### 3.3.4 Data Manipulation

- Using dplyr functions: filter(), select(), mutate(), group_by(), summarise().
- Using tidyr for reshaping data: pivot_longer(), pivot_wider().

## 3.4 Model Evaluation Techniques

- Performance metrics: Accuracy, Precision, Recall, F1-score.
- Cross-validation using train() from the caret package.

# CHAPTER 2

# PROJECT METHODOLOGY

## 2.1 PROPOSED WORK

The methodology for this project involves a systematic approach to analyzing student exam scores using data science techniques in R. The proposed work is divided into the following phases:

### 1. Data Collection and Preprocessing

- Import the dataset containing student exam scores and related attributes (e.g., study time, attendance, gender, etc.).
- Handle missing values, remove duplicates, and normalize the data as needed.
- Convert categorical variables into appropriate numerical formats using encoding techniques.

### 2. Descriptive Statistical Analysis

- Perform summary statistics to understand the distribution, central tendency, and spread of exam scores.
- Use visual tools such as histograms, boxplots, and correlation matrices to identify trends and relationships.

### 3. K-Means Clustering

- Apply K-Means clustering to segment students into distinct performance groups based on exam scores and relevant features.
- Determine the optimal number of clusters using the Elbow Method or Silhouette Score.

- Visualize clusters for better interpretation.

## 4. Regression Modeling

- Build a regression model (e.g., linear regression) to identify and quantify the impact of different factors on exam scores.
- Evaluate the model using metrics such as $R^2$, RMSE, and residual analysis.

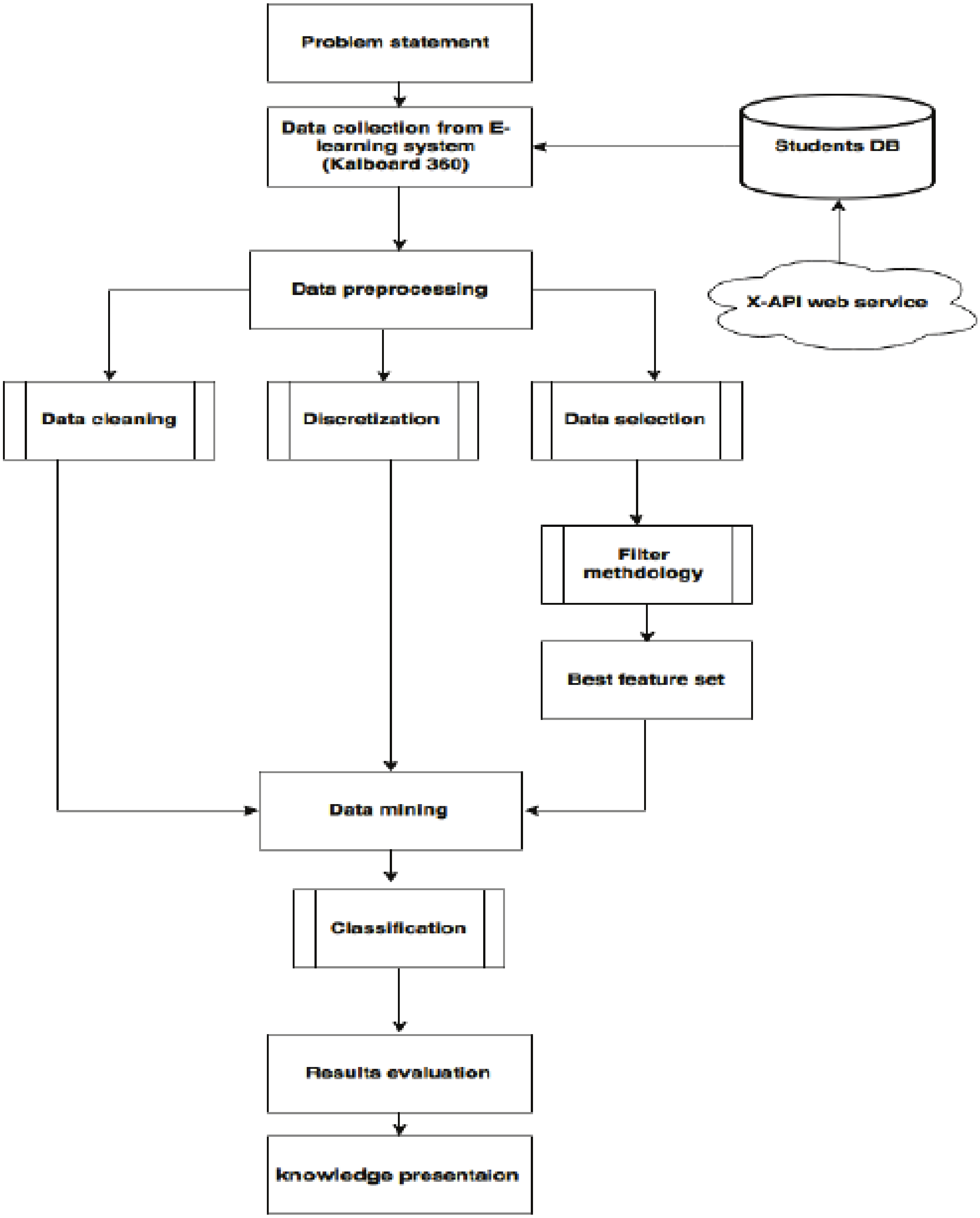
## 5. Classification Analysis

- Categorize students into performance levels (e.g., high, average, low) based on exam scores using classification algorithms such as Decision Trees or Logistic Regression.
- Split the dataset into training and testing sets, and evaluate the model using accuracy, precision, recall, and F1-score.

## 6. Result Interpretation and Insight Generation

- Compare performance across different models and techniques.
- Draw insights and provide recommendations for educators to enhance student learning outcomes.

## 2.2 BLOCK DIAGRAM



# CHAPTER 3

# MODULE DESCRIPTION

## 4.1 DATA INPUT AND PREPROCESSING MODULE

Accepts exam scores from different sources (CSV file, database, manual entry).

Cleans and organizes data for analysis

## 4.2 STATISTICAL COMPUTATIONAL MODULE

Computes Mean, Median, and Mode.

## 4.3 DATA VISUALIZATION AND ML USED MODULE

Generates histograms, boxplots, and bar charts for better understanding.

Predicts future student performance trends.

## 4.4 REPORT GENERATION MODULE

Generates summary reports with findings and recommendations.

# CHAPTER 4

# CONCLUSION & FUTURE SCOPE

## 4.1 CONCLUSION

In analysis of exam scores using R has provided valuable insights into student performance through both statistical and machine learning techniques. Descriptive analysis helped in understanding the distribution and variability of scores, offering a foundational view of the dataset. K-Means clustering effectively grouped students into performance-based clusters, revealing hidden patterns among different learner groups. Regression modelling identified key factors influencing academic outcomes and enabled score prediction with reasonable accuracy. Classification techniques further enhanced the analysis by categorizing students into performance levels, aiding in targeted interventions. Overall the integration of these methods demonstrated how data-driven approaches can support decision-making in education. The project highlights the importance of leveraging analytics for monitoring academic progress, identifying at-risk students, and improving learning outcomes. Future work can include real-time data analysis, feature expansion (like socio-economic factors), and deployment of predictive models in academic management systems.

## 4.2 FUTURE SCOPE

This study provides a strong foundation for further exploration in the field of educational data analytics. In the future, the analysis can be enhanced by incorporating additional variables such as attendance records, socio-economic background, parental education levels, and learning behavior to improve prediction accuracy and depth of insight. Time-series analysis could be used to track student performance trends over multiple academic terms, enabling early identification of at-risk students. Additionally, the implementation of real-time dashboards using tools like R Shiny could help educators and administrators monitor performance dynamically and take timely actions.

# APPENDIX A – SOURCE CODE

```
library(ggplot2)
library(dplyr)
library(caret)
library(cluster)
library(factoextra)
library(e1071)
  # Load dataset
  data <- read.csv("C:/Users/padma/Downloads/results.csv")
  # Descriptive statistics
  summary(data)
  sapply(data[,2:8], sd)  # standard deviation of scores
  # Correlation matrix
  cor_matrix <- cor(data[,2:7])
  print(cor_matrix)
  # K-Means Clustering
  scores <- scale(data[,2:7])  # Standardize
  kmeans_model <- kmeans(scores, centers = 3)
  data$Cluster <- as.factor(kmeans_model$cluster)
  # View first few cluster assignments
```

```
head(data[, c("Total", "Cluster")])
# Visualize clusters
fviz_cluster(kmeans_model, data = scores)
# Regression model (predicting Total)
reg_model <- lm(Total ~ Hindi + English + Science + Maths + History + Geograpgy,
data = data)
summary(reg_model)
# Add predicted total to dataset
data$Predicted_Total <- predict(reg_model, data)
# Classification (Results: Pass/Fail)
data$Results <- as.factor(data$Results)
set.seed(123)
trainIndex <- createDataPartition(data$Results, p = 0.7, list = FALSE)
trainData <- data[trainIndex, ]
testData <- data[-trainIndex, ]
model <- train(Results ~ Hindi + English + Science + Maths + History + Geograpgy,
data = trainData, method = "glm", family = "binomial")
# Predict and evaluate
pred <- predict(model, newdata = testData)
confusionMatrix(pred, testData$Results)
```

# APPENDIX B – SCREENSHOTS

**summary(data)**

```
       X             Hindi          English         Science         Maths      
 Min.   :  0.0   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00  
 1st Qu.:249.8   1st Qu.:26.00   1st Qu.:26.00   1st Qu.:25.00   1st Qu.:25.75  
 Median :499.5    Median :53.00    Median :50.50    Median :50.00    Median
:49.00  
 Mean   :499.5   Mean   :51.65   Mean   :50.11   Mean   :49.44   Mean   :49.55  
 3rd Qu.:749.2  3rd Qu.:77.00  3rd Qu.:75.00  3rd Qu.:73.25  3rd Qu.:74.00  
 Max.   :999.0   Max.   :99.00   Max.   :99.00   Max.   :99.00   Max.   :99.00  
    History        Geograpgy         Total          Results         Div       
 Min.   : 0.00   Min.   : 0.00   Min.   :103.0   Min.   :0.00   Min.   :0.000  
 1st Qu.:24.00   1st Qu.:26.00   1st Qu.:254.0   1st Qu.:0.00   1st Qu.:1.000  
 Median :49.00    Median :49.00    Median :296.0    Median :0.00    Median
:2.000  
 Mean   :49.03   Mean   :50.03   Mean   :299.8   Mean   :0.35   Mean   :1.856  
 3rd Qu.:73.25  3rd Qu.:75.00  3rd Qu.:349.2  3rd Qu.:1.00  3rd Qu.:3.000  
 Max.   :99.00   Max.   :99.00   Max.   :505.0   Max.   :1.00   Max.   :3.000  
```

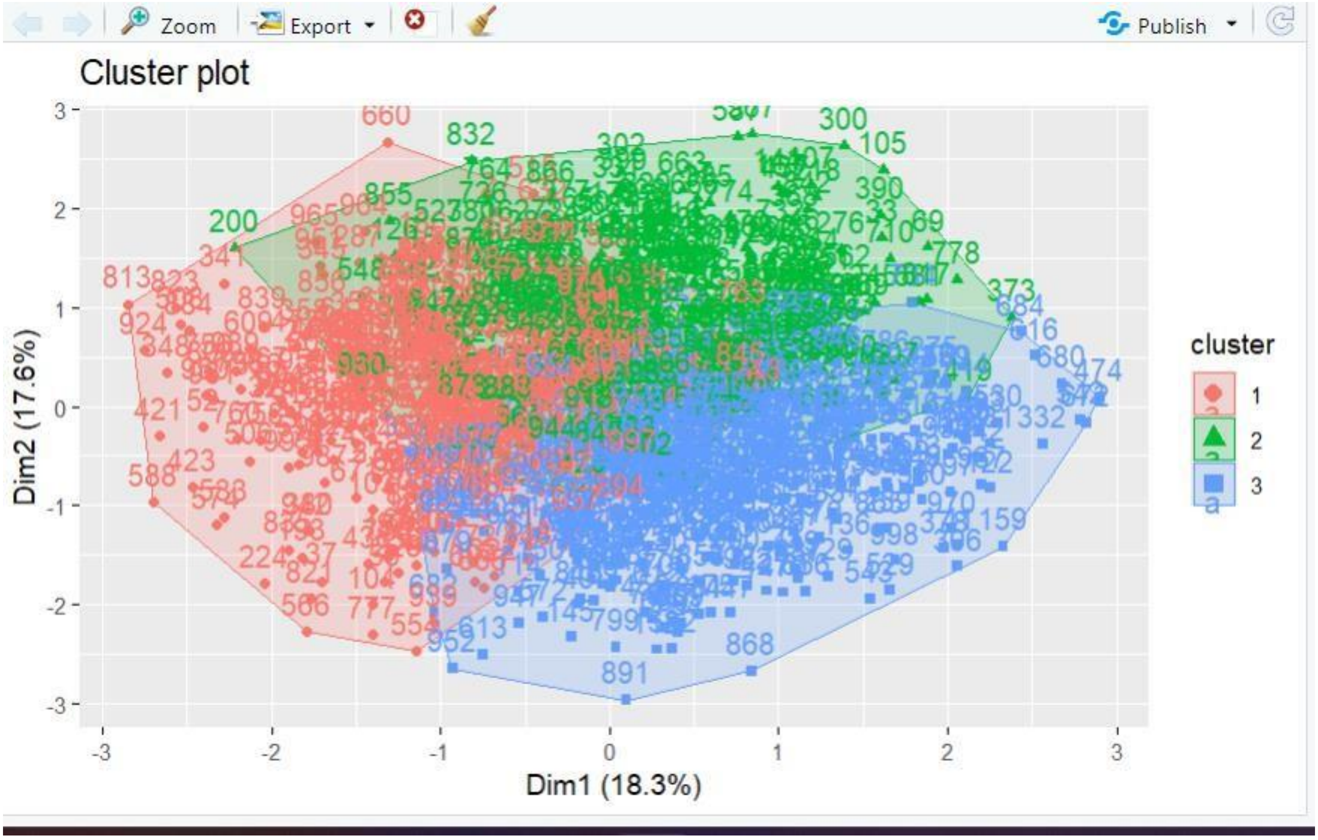
**Standard deviation of scores**

```
    Hindi   English   Science     Maths   History Geograpgy     Total 
 29.47191  28.04850  28.92111  28.63245  28.76975  28.71027  71.86524 
```

**Correlation matrix**

```
               Hindi      English      Science       Maths      History    Geograpgy
Hindi  1.00000000  0.036454895  0.037615847  0.041239550 -0.024652488 -
0.064189091
English         0.03645489      1.000000000    -0.022410890    -0.011268721    -
0.001812014  0.052252624
Science          0.03761585    -0.022410890       1.000000000       0.056313623
0.005605054  0.028845817
Maths              0.04123955    -0.011268721       0.056313623       1.000000000
0.011524990 -0.001140353
History          -0.02465249    -0.001812014       0.005605054       0.011524990
1.000000000 -0.022318004
```

# DATA VISUALIZATION



# REFERENCES

1. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.
3. Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.
4. Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.
5. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest.*RNews*,2(3),18–22