**Introduction**

The objective of this project is to examine dataset using R language analytical tools and techniques. So, we have selected an HR dataset of a pharmaceutical company to conduct the analysis. Understanding HR data becomes critical for organizational success in the business, where people acquisition, retention, and development play crucial roles. The dataset used in this analysis includes employee-related data from internal databases, including training records, performance indicators, and demographics.

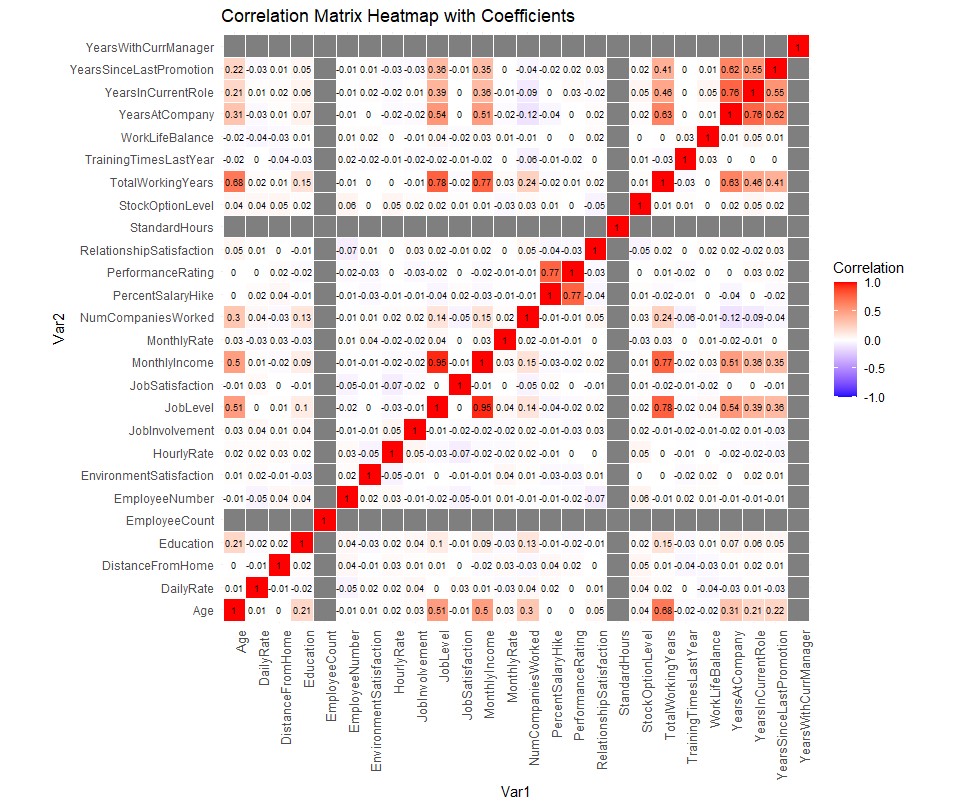
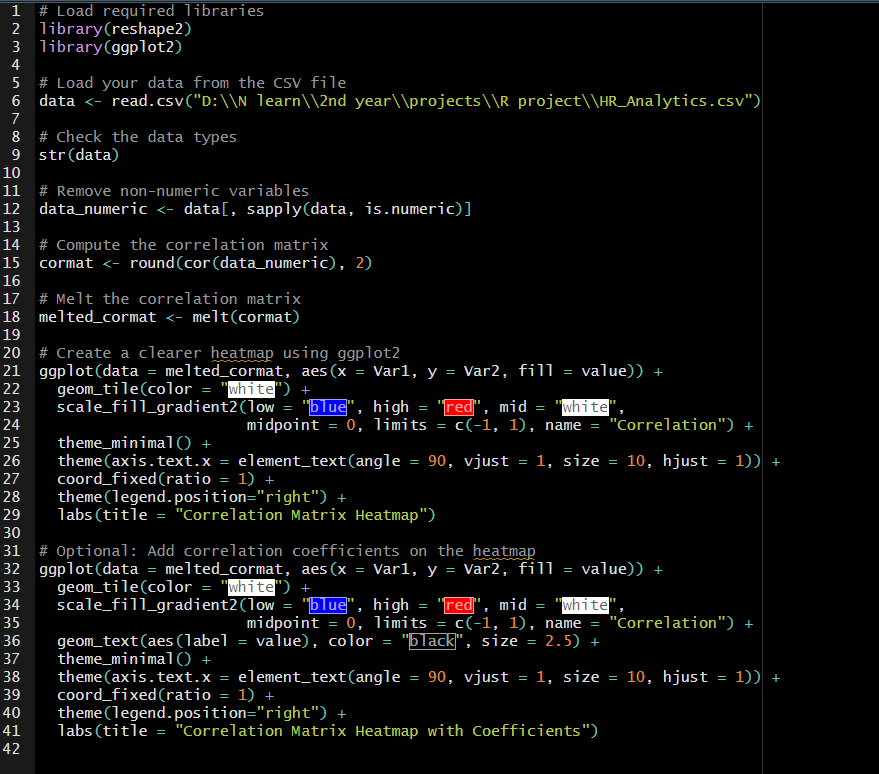
Finding insights that help improve decision-making, optimize HR initiatives, and lead to a more productive and efficient staff are some of our goals. This project uses R's exploratory data analysis and predictive modelling tools to deliver useful insights and recommendations for HR management.

We conducted our project by dividing our dataset into 5 main categories,

1. Age
2. Gender
3. Experience
4. Education
5. Marital Status

Under these categories we conducted an analysis to find valuable insights about how these 5 categories impact a person who works in a company.

To find which columns in our dataset impact these categories the most, we used a correlation matrix heatmap. With this we were able to get an initial idea about our dataset.



1. **Age**

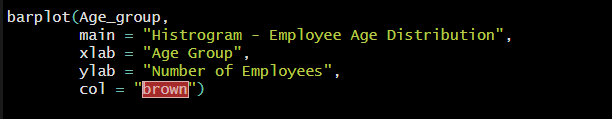
This HR dataset analysis's age part is a crucial component for understanding how age affects different components of the workforce.

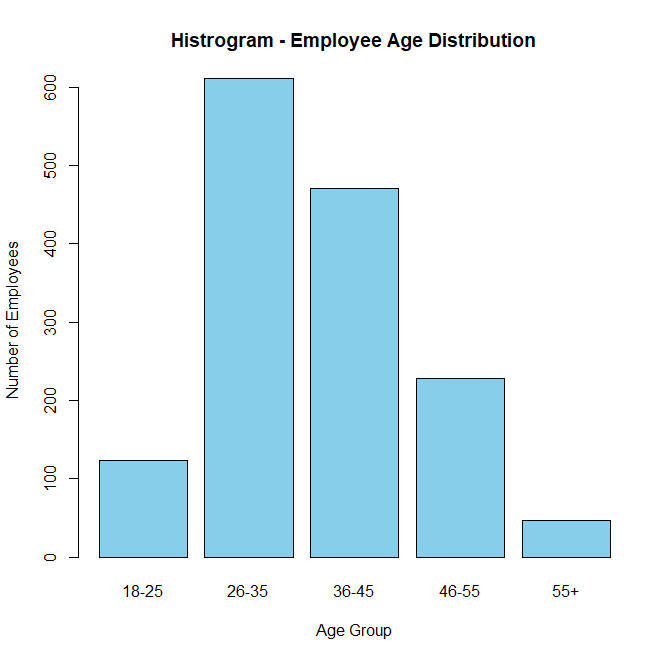
We hope to identify patterns, trends, and potential biases related to various age groups by analyzing this data.

**Data Visualization**

In this analysis of age, the mainly used plots are barplots and boxplots. In this category the data type that commonly occurred is categorical data. barplots and boxplots are R functions, typically used to visualize categorical data.

Shown below is the code used to visualize the age distribution of this company.

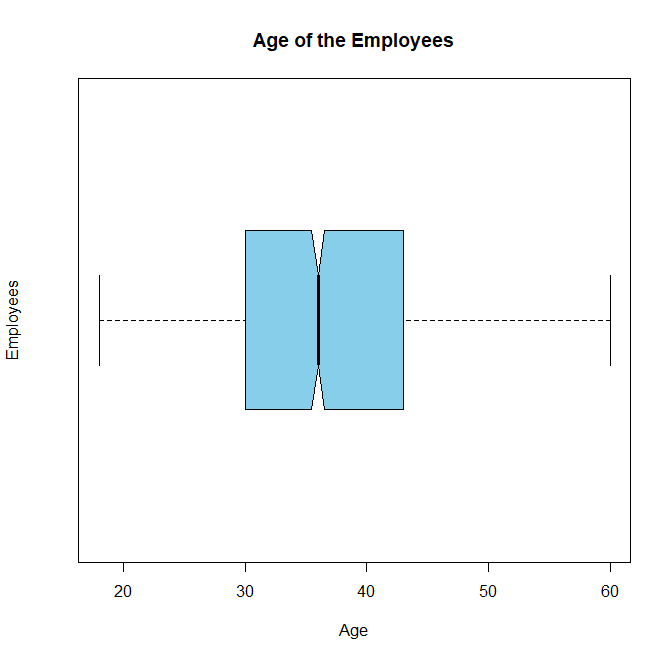




Boxplot for the age distribution of this company.

A screen shot of a computer

Description automatically generated

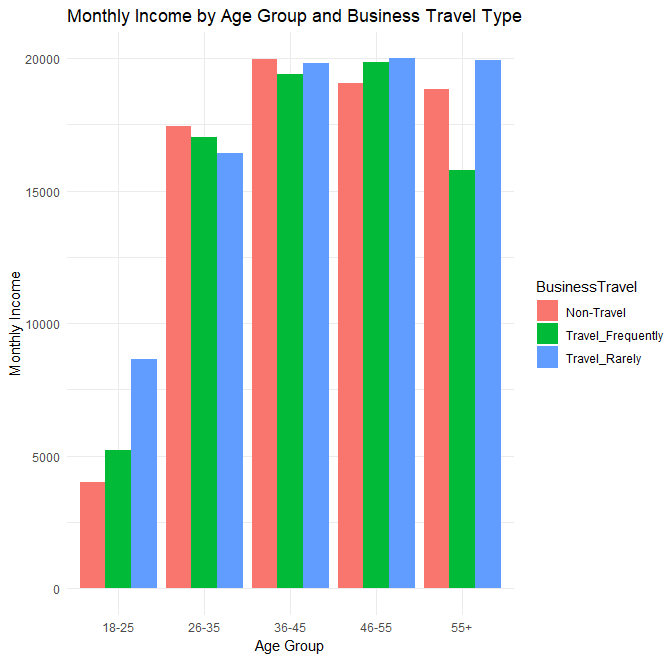
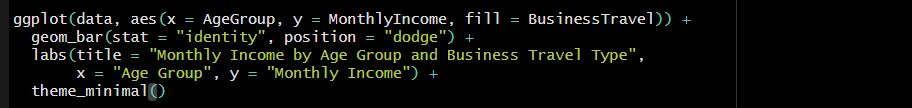


By this we can clearly say that in this company most employees are middle-aged people. With the average age distribution around 35 – 38.

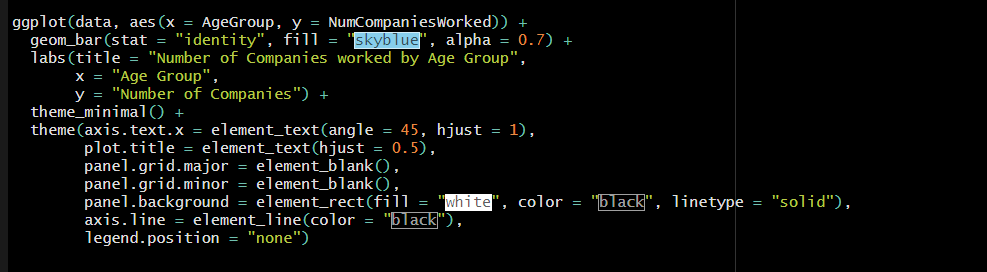
We can also investigate in this dataset the relationship between an employee's age group, the amount of business travel they do, and how it affects their pay each month.

Group barplots were used to visualize this data representation.

We observed that the dataset's business travel section had two data entries that were comparable to each other: "Travel\_Rarely" and "TravelRarely." We have changed any instances of "TravelRarely" to "Travel\_Rarely" to maintain consistency.

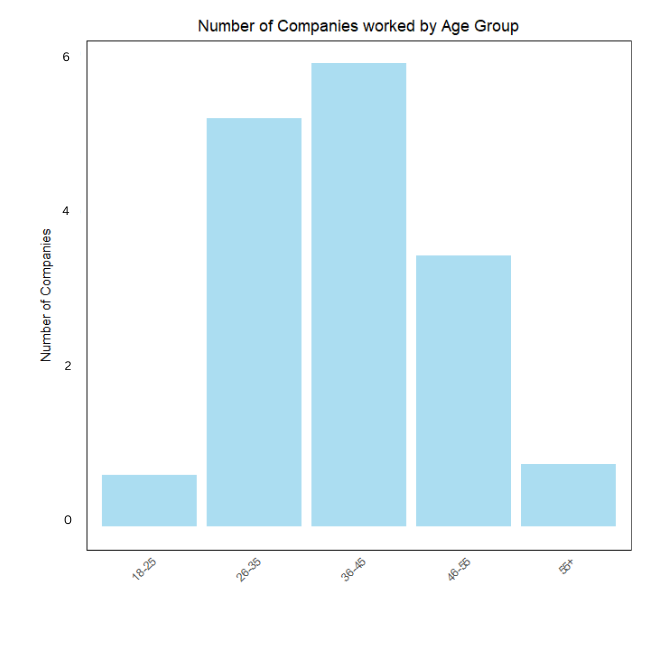


With this we can say that in most of the age groups the business travel frequencies are almost equally divided. And the employees in the age group of 18 – 25 get less Income compared to the other employees.

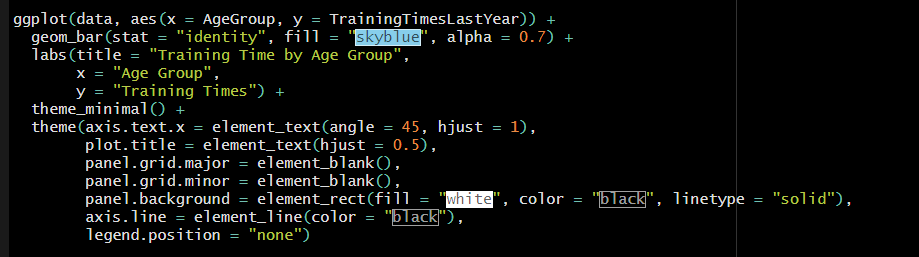


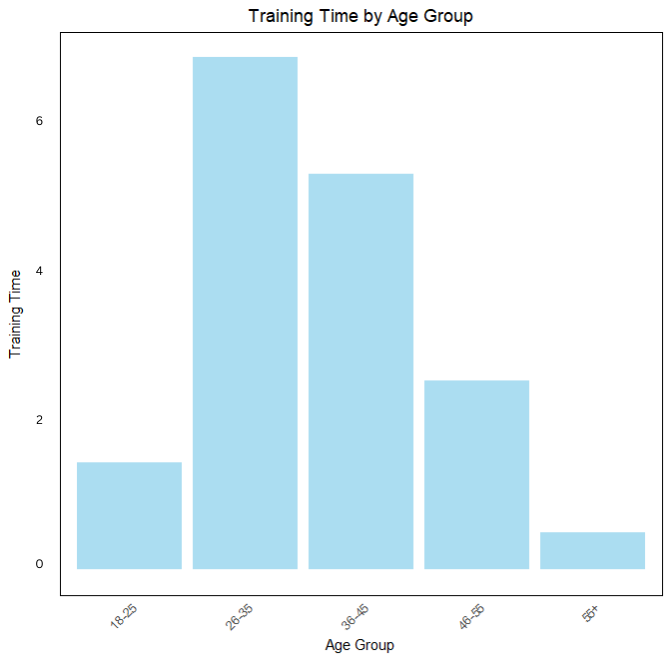
From the above shown code we visualize the data which states the number of companies that the employees worked in according to their age group.

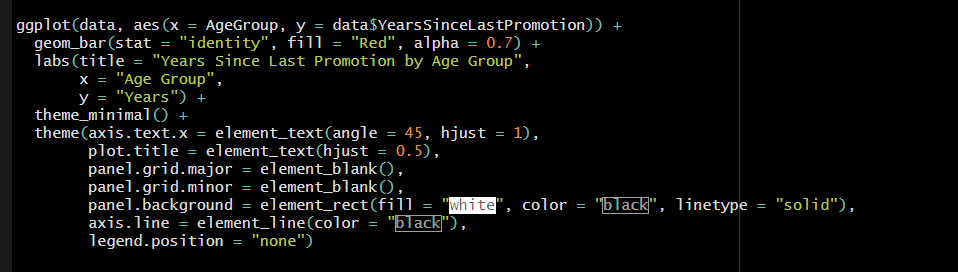
To represent the data, we have used a barplot using ggplot libraries.



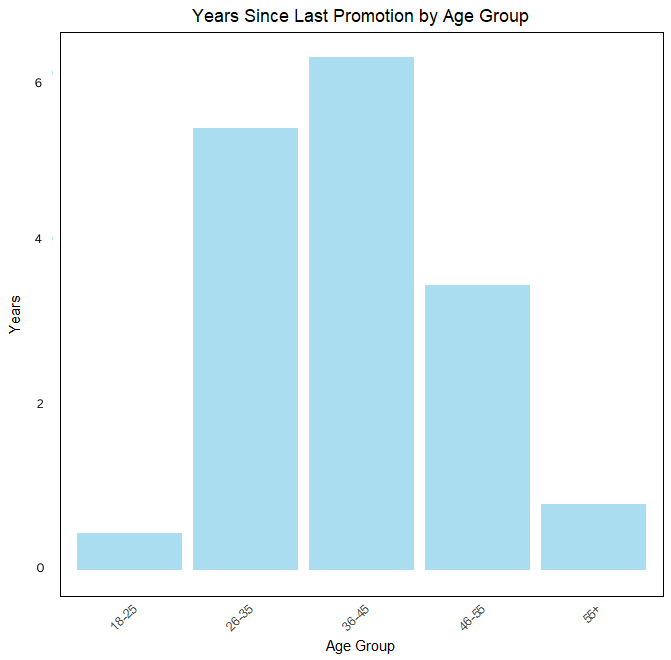
From the plot we can say that the employees in the age group 25 – 35 , 36 – 45, 46 – 55 have worked in many companies compared to other age groups.

Shown below is the code that was used to visualize this data.We used ggplot libraries to plot a barplot to visualize this data.





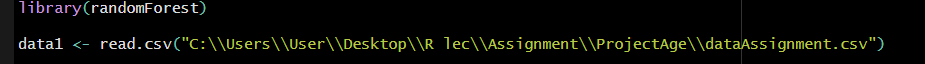
The above code is used to plot a barplot using ggplot libraries to visualize how the time since the last promotion employees had change according to their age group.

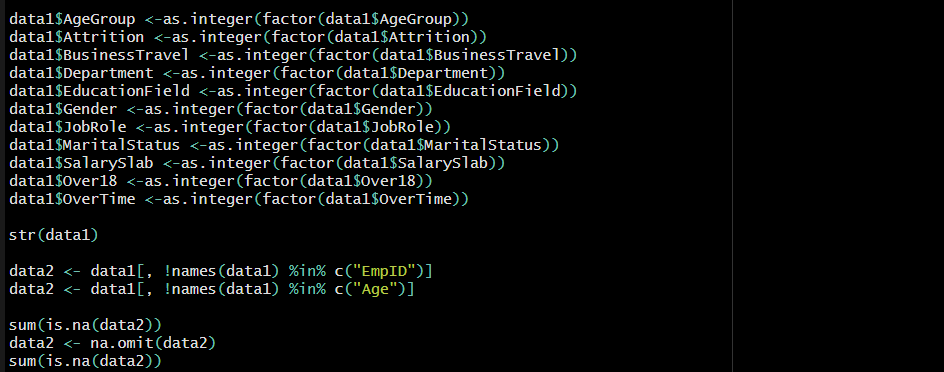


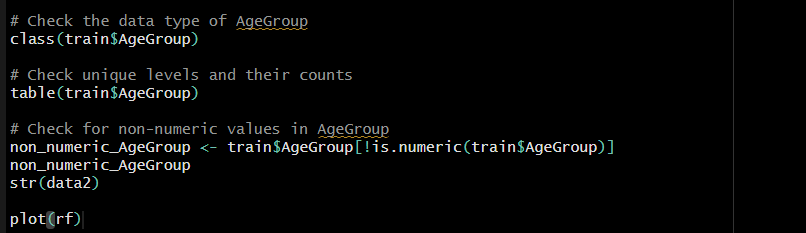
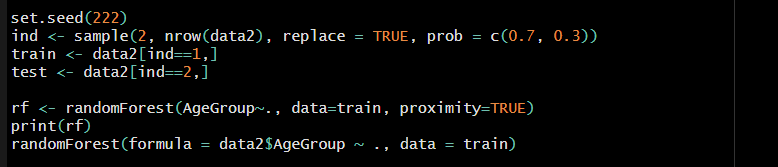
**Data analytics through models**

**Random Forrest Classification**

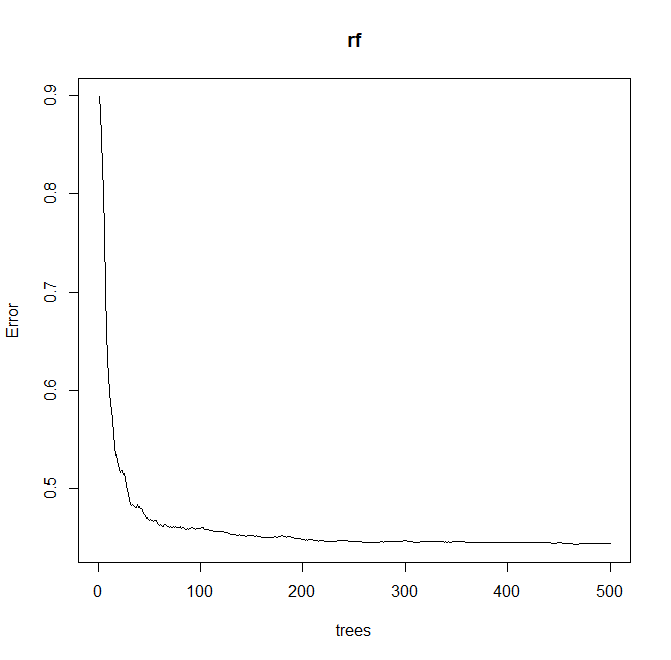
Random Forrest classification model helps us to inspect the dataset structure, understand the distribution of the target variable, check, and visualize aspects of the model's behavior or performance. This helps in understanding and evaluating the model and the data before using it for predictions or further analysis.



After loading the needed libraries and dataset data cleaning and preprocessing was done.

** **

With this we can plot an Error Rate Plot.

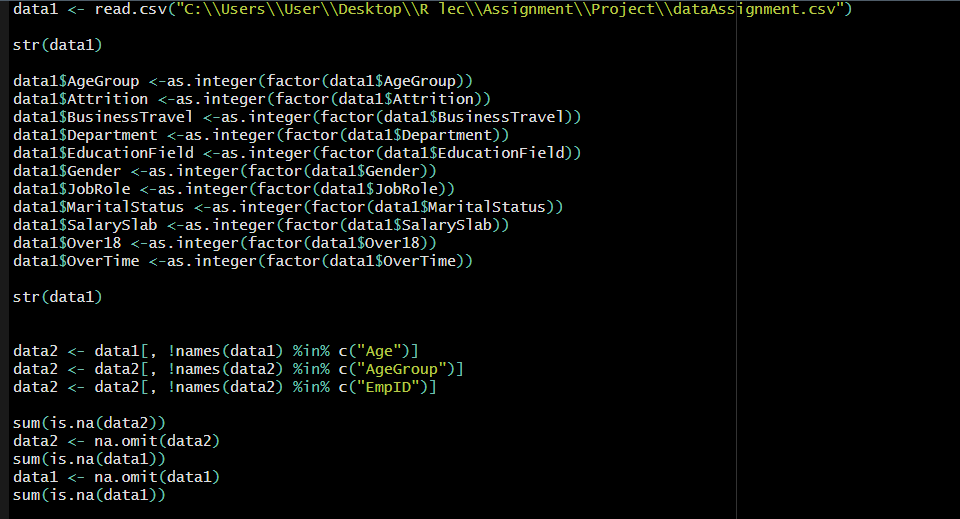
****

This illustrates how more trees are added to an ensemble learning process, which tends to increase forecast accuracy by lowering mistakes.

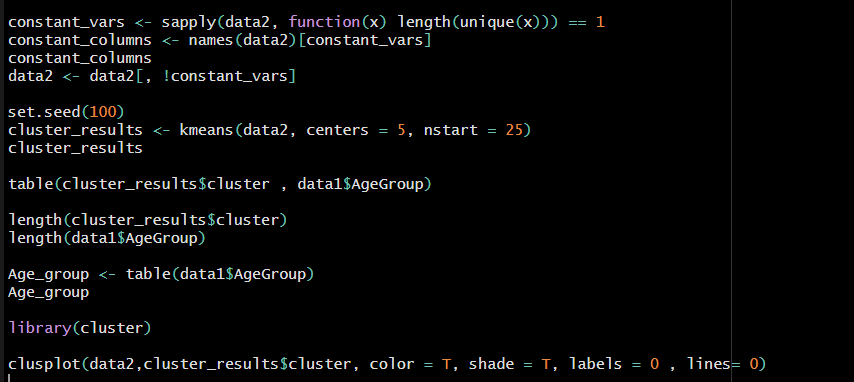
An essential tool for comprehending how the predicted accuracy of the Random Forest model changes with tree count is an Error Rate graphic.

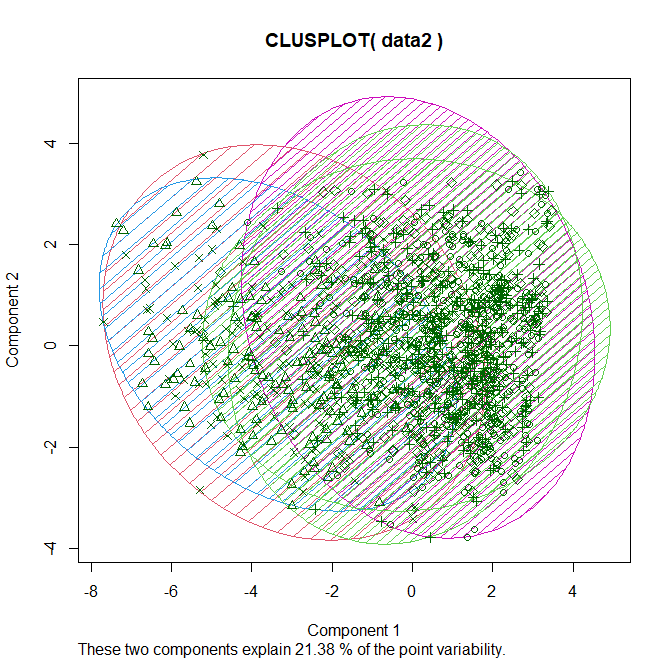
**Clustering Model**

In machine learning, a sort of unsupervised learning technique called a clustering model groups or clusters data according to similarities between them. It will help us to identify which data points are similar.

****

First, we cleaned and preprocessed the dataset to make it properly fit into the model.

cluster library was used.



Although the variability percentage is low from this, we can say that age group of the employees can be identify from the model.

**Findings and Discussion**

From visualization part there was a clear trend on the years since last promotion and training times. From that we can say that if an employee gets less training time that employee is going to get promoted vice versa. And from the plots of the training times and number of companies worked we can get that if an employee work in various companies that employee is going to get more training inside the company. And it is possible to conclude that employees who travel for work earn more per month than workers who do not travel for business. All age groups in the dataset appear to follow this pattern.

In conclusion, a thorough examination of the dataset's age-related features was made possible by the combination of modeling and data visualization techniques, which provided useful insights and laid an outline for future in-depth research in the relevant field.

1. **Gender**
2. **Experience**
3. **Education**
4. **Marital Status**