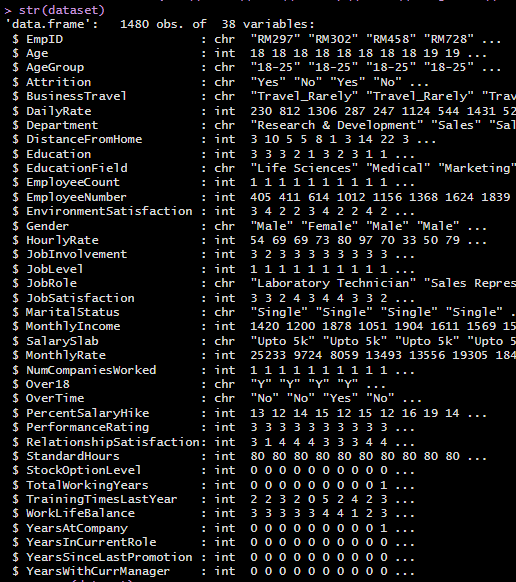
**Introduction**

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.



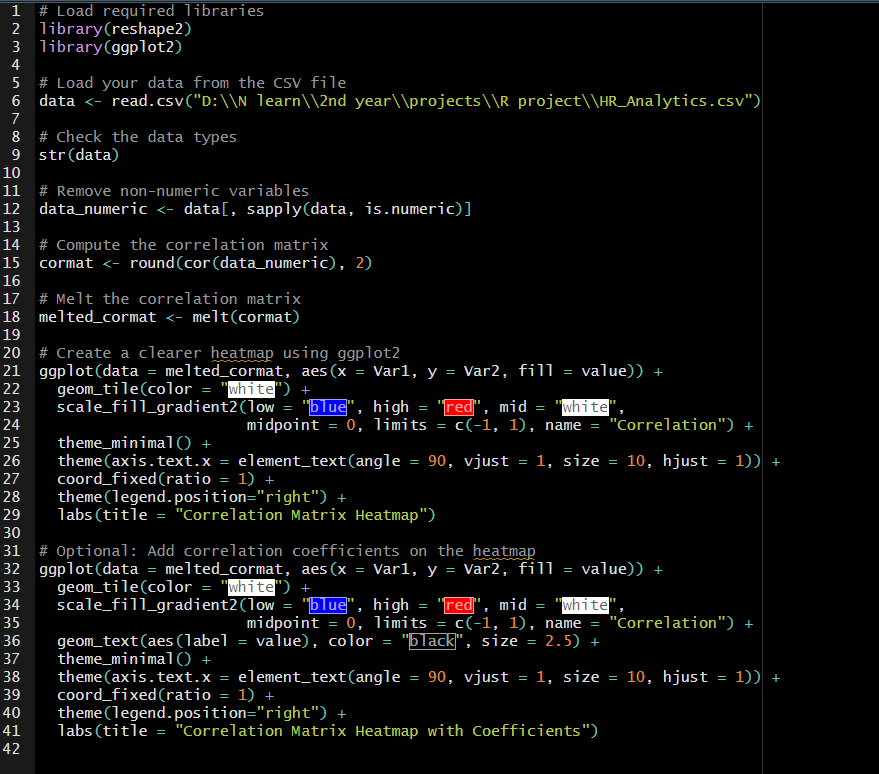
The structure of the HR dataset we selected is shown above.

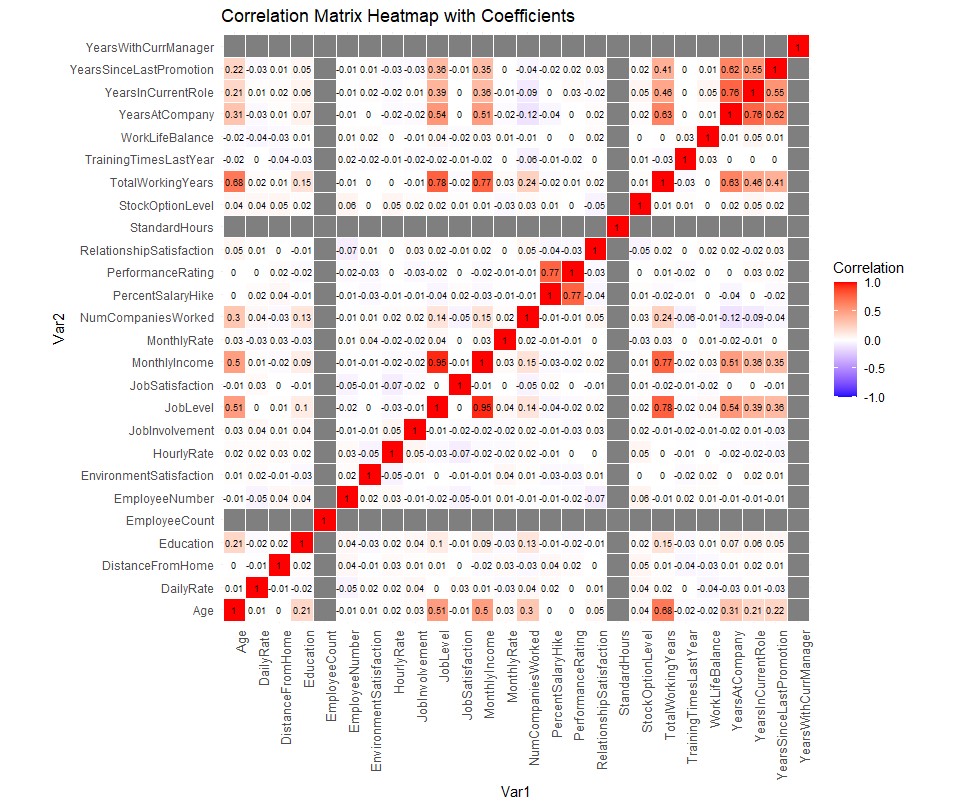
We conducted our project by dividing our dataset into 5 main categories,

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**

Analyzing age in HR datasets provides crucial insights for understanding workforce demographics, fostering an inclusive workplace and more.

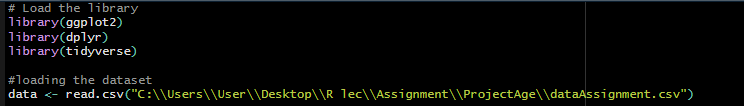
This analysis has been performed on the HR dataset we chose to identify the hidden information inside the dataset.

**Data Visualization**

In this analysis , we have used several types of plots to perform data visualizations. From this we can get hidden information on age that we cannot by looking at the dataset directly.

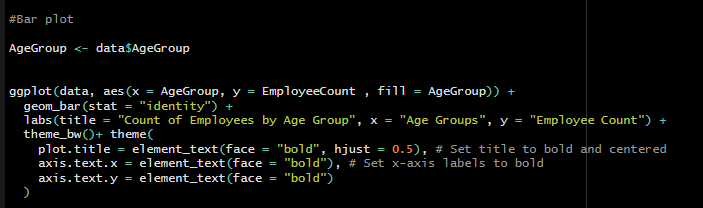
The graphs and the visualizations used in this analysis were accessed by using the ggplot2 libraries.

First, we need to import our data set into the R script and install the necessary libraries.



After loading and installing the necessary data and the libraries we looked at our dataset and did some research using the heat map used to get the correlation with each column of the dataset. From that we have chosen some data columns from the data set to get insights with the age.

We have used a barplot to visualize the employee count by age group. Shown below is the R code we used to plot this barplot.



A graph of a number of people

Description automatically generated

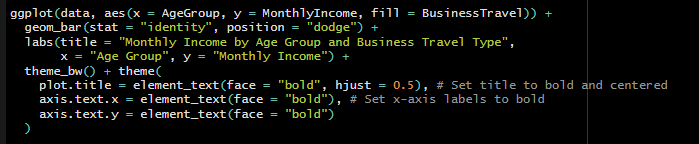
From this plot we can say that most employees in this company are in the age group of 26-35. And this plot has a bell type shape so we can say that this company’s’ age distribution is a standard distribution and doing statistical calculation will give us a higher accuracy.

**Relationship between an employee's age group, the amount of business travel they do, and how it affects their pay each month.**

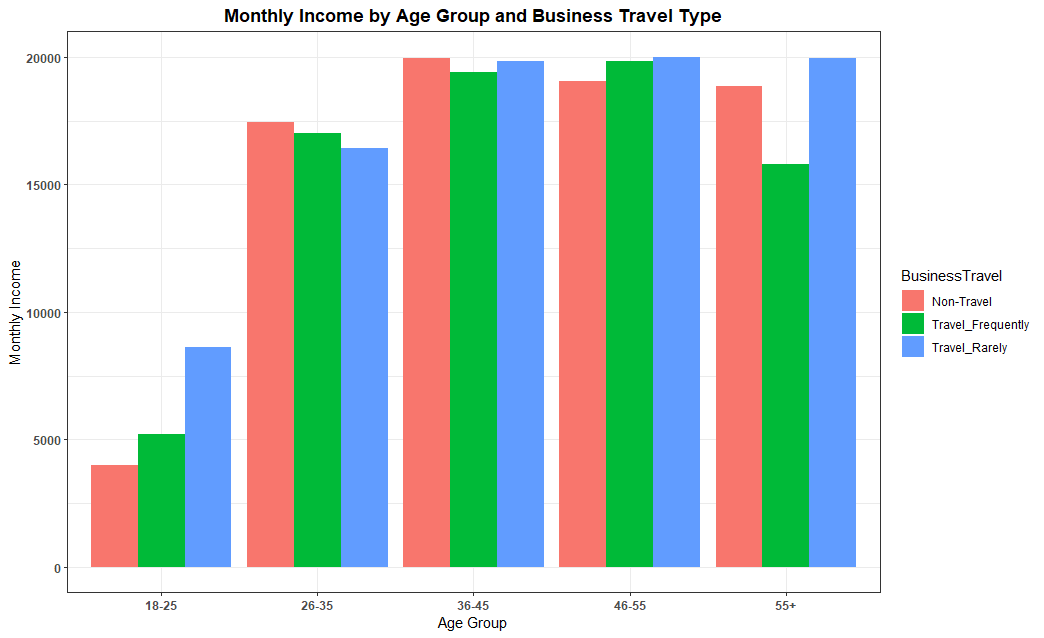
Group barplots were used to visualize this data representation.

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.

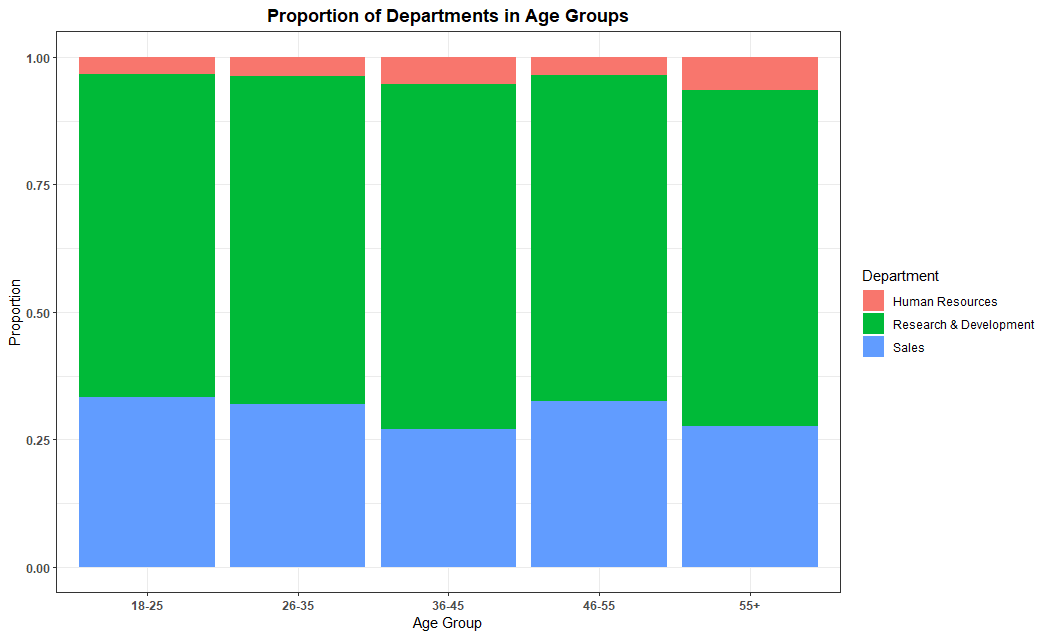




Shown below is the plot we generated from the R code shown above.



**Proportion of Departments in Age Group.**

****

With this mosaic plot we can get a better understanding of how the employees in this company have been distributed according to the departments they work in.

So, from this we can say that this company’s focus is on research and development.

A computer code with text on it

Description automatically generated

Shown above is the R code used to generate the mosaic plot.

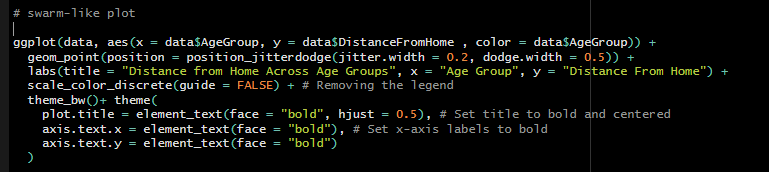
**Distance from home Across Age Groups**

**A graph of a number of people

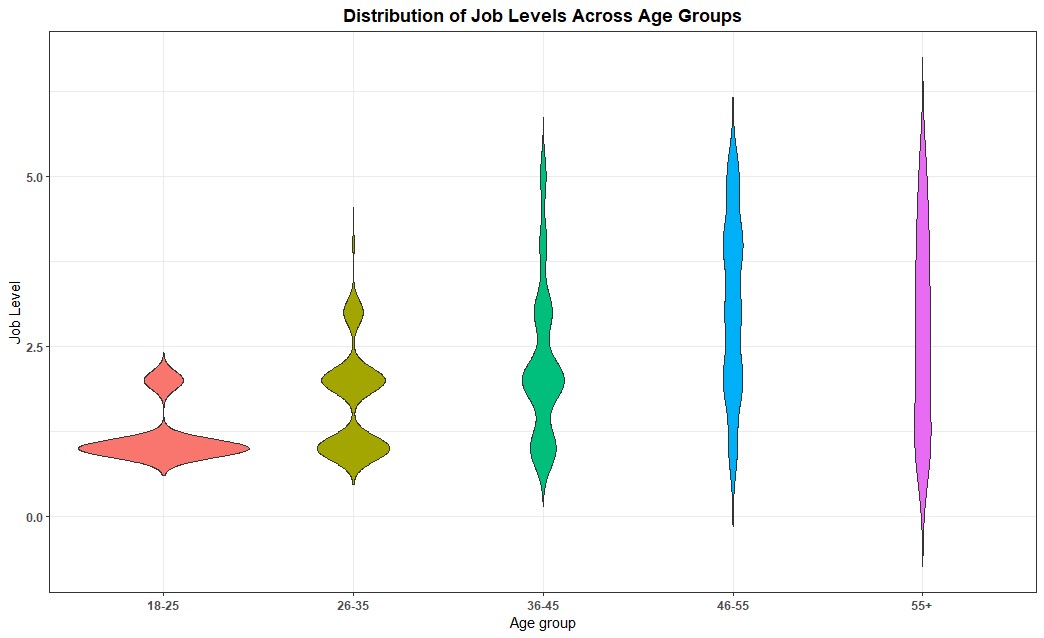
Description automatically generated with medium confidence**

From this plot we can see almost all the employees of this company live near to the company premises.

Given below is the R code used to generate this plot.



**Distribution of Job Levels Across Age Group**



This is a violin plot representing the distribution of the job level across age groups. As you can see from this plot the job level gradually increases as the age increases.

Given below is the r code used to generate this plot.

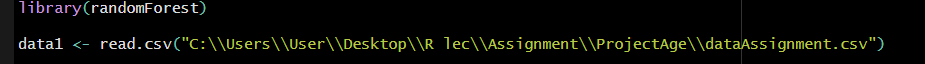
A screen shot of a computer code

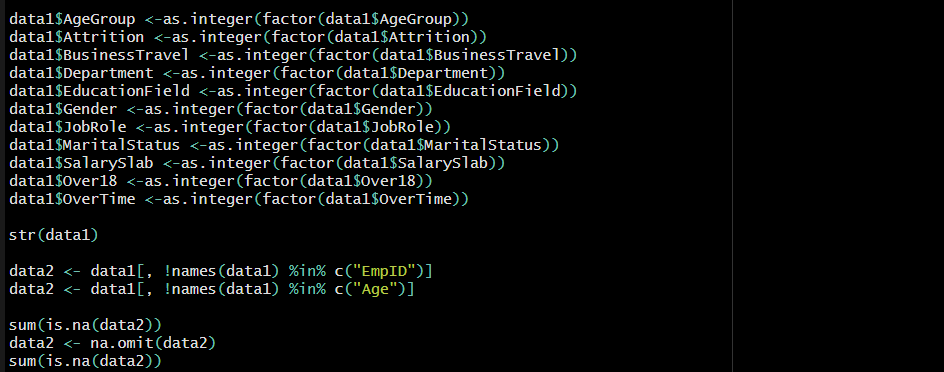
Description automatically generated

**Data analytics through models**

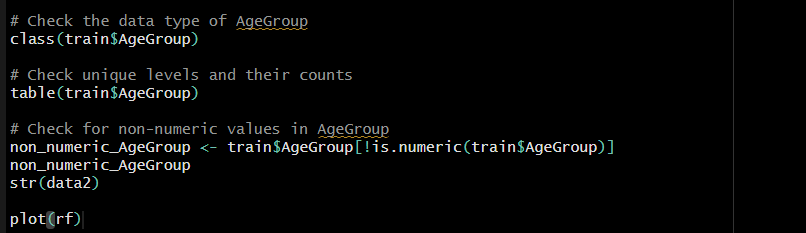
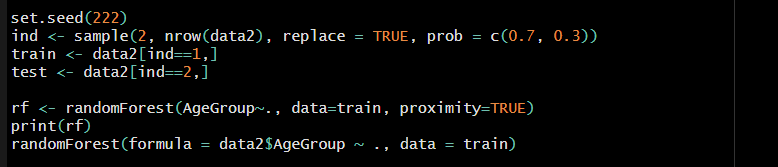
**Random Forrest Classification on Age group**

Using a Random Forest model to forecast age groups in an HR dataset can help create a more inclusive and productive work environment, provide insightful information about the demographics of the workforce, and assist in developing focused HR campaigns.

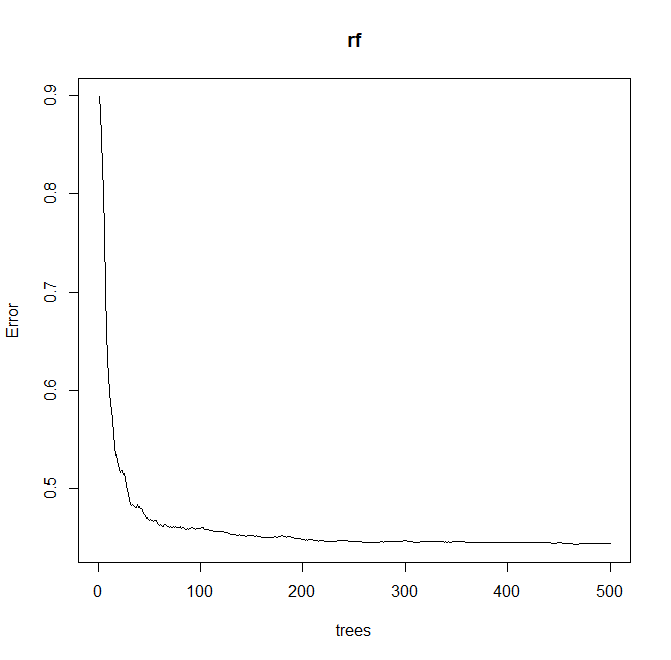




After loading the needed libraries and dataset data cleaning and preprocessing was done.

** **

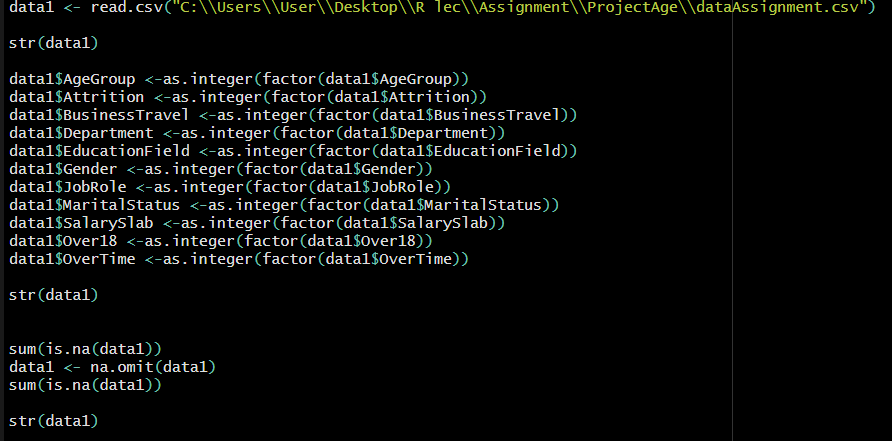
With this we can plot an Error Rate Plot.

****

By evaluating this error rates, we can assess the overall performance and stability of the random tree. From the above tree the error rate has decreased meaning the model has been stabilized after 300 trees.

**Naïve Bayes Classification on Age Group.**

The Bayes theorem, which determines the probability of a label (class) given the observed features, is the foundation of naive Bayes classifiers.

****

A computer screen with white and orange text

Description automatically generated



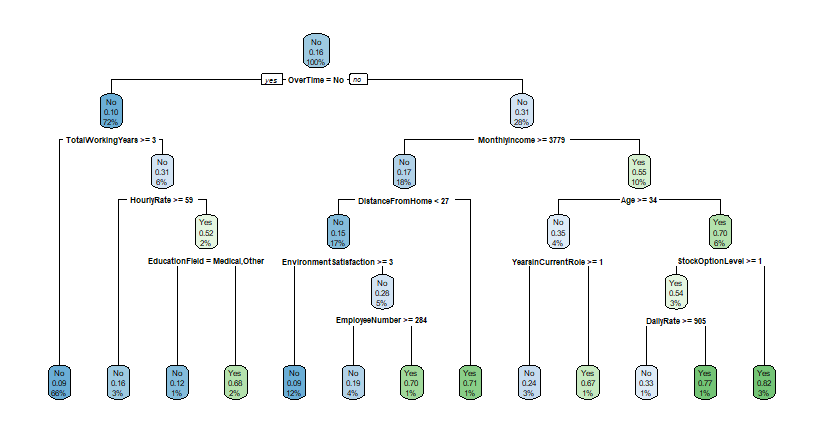
In this classification model we were able to classify the age group into groups with an accuracy of 0.7781.

A screenshot of a computer

Description automatically generated

**Decision tree classification on Attrition**

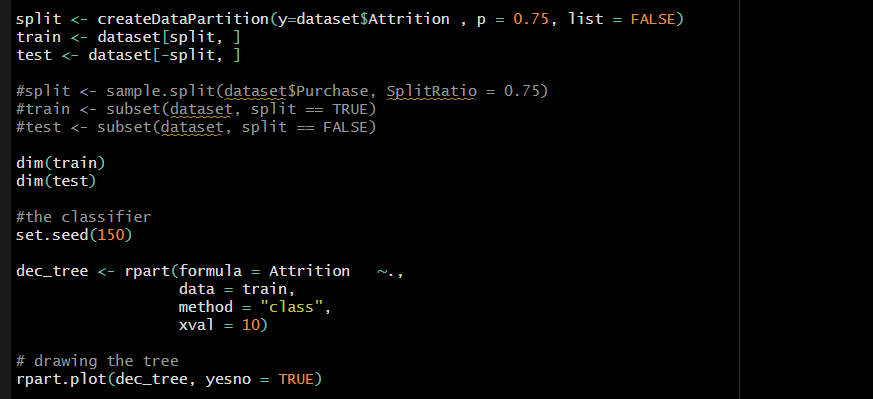
Decision tree classification is a useful tool in many applications because it provides comprehensible results and a plethora of insights into data patterns, feature relevance, and predictive capabilities.

****

After evaluating this model though, a confusion matrix has an accuracy of 0.7967.

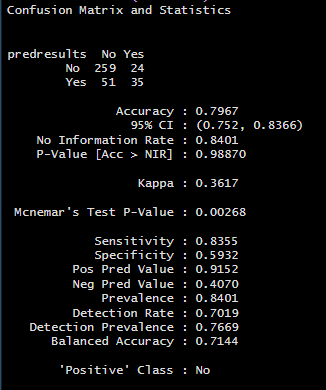
A computer screen with text on it

Description automatically generated

****

**A computer screen with white text

Description automatically generated**

****

In this model removing the Joblevel from the dataset gives us an increased accuracy for this dataset.

**Findings and Discussion**

From the employee count bar plot we can derive that most of the employees work in this are in the age group 26-35 or above so when the company give benefits to the employees’ customization of benefits and programs that cater to the specific needs according to the nature of the employees will more appreciated. As an example, giving more weight to retirement planning or healthcare benefits will be more appropriate because most of the employees are in the older age groups.

From the plot that shows how the distance from home across the age group we can say that most of the employees who works in this company have a distance between 10 – 20 Kms. So, arranging a daily commute for them will be more beneficial because it will increase employee satisfaction, enhance productivity, improve punctuality and the attendance of the employees.

When we look at the violine plot demonstrates distribution of job levels across age group we can say that the youngest and oldest age groups have the widest distribution in job levels. As a result, there is a greater variety of jobs among these age groups, with some members of these groups having extremely high or extremely low job levels.

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**