**INTRODUCTION**

In this project, the IBM HR Employee Attrition and Performance dataset was used to analyze employee attrition. The dataset includes various features such as employee demographics, job satisfaction, performance ratings, and attrition status. Variables in the dataset include "Age", "Attrition", "BusinessTravel", "Department", "DistanceFromHome", among others.

Visualizations and logistic regression analysis were performed using these variables. Visualizations were employed to understand relationships and patterns within the data, while logistic regression analysis was applied to predict the likelihood of employee attrition.

The results have helped identify factors influencing attrition rates and understand their impact on the likelihood of employees leaving the company. This analysis contributes to the development of strategies by human resources departments to enhance employee engagement and reduce attrition rates.

**Exploratory Data Analysis(EDA)**

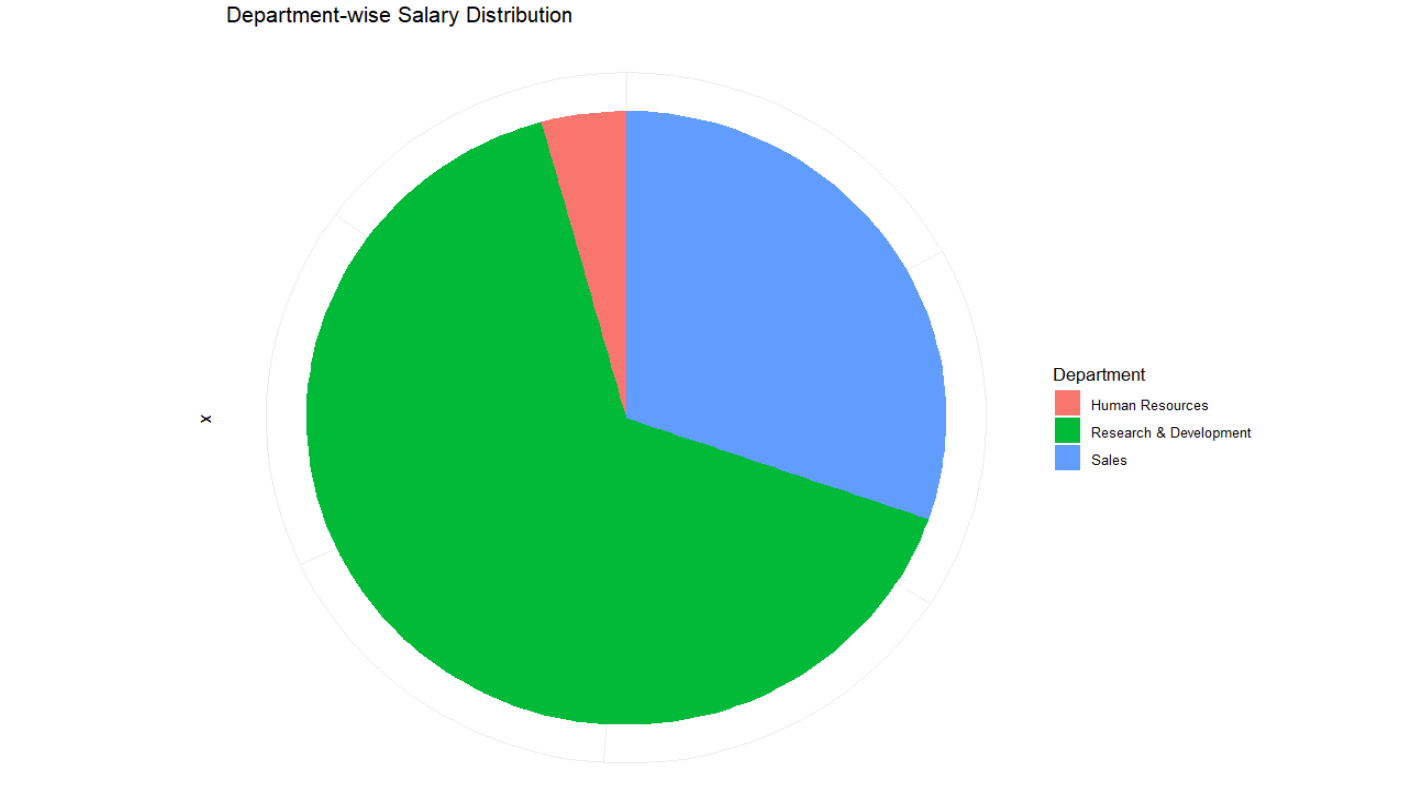
In this project, Exploratory Data Analysis (EDA) was conducted on IBM's HR Employee Attrition and Performance dataset, which contains 35 variables and 1,470 observations. Initially, categorical variables that could be converted to factor structures were identified and appropriately transformed. A check for missing data was performed, and it was found that there were no missing values. To better understand the distribution of the data, the table() function was used to examine the frequencies of categorical variables.

To gain a deeper understanding of the data, various visualizations were created using the ggplot2 package. For example, histograms were used to examine the age distribution of employees, box plots were used to explore the relationship between job satisfaction and attrition status, and scatter plots were used to analyze relationships between numerical variables. These visualizations are crucial for understanding the structure of the data and preparing for the modeling process.

In conclusion, the EDA process has helped us understand the fundamental characteristics and relationships within the dataset, confirming the absence of missing data and anomalies. This analysis has provided a solid foundation for subsequent modeling steps.

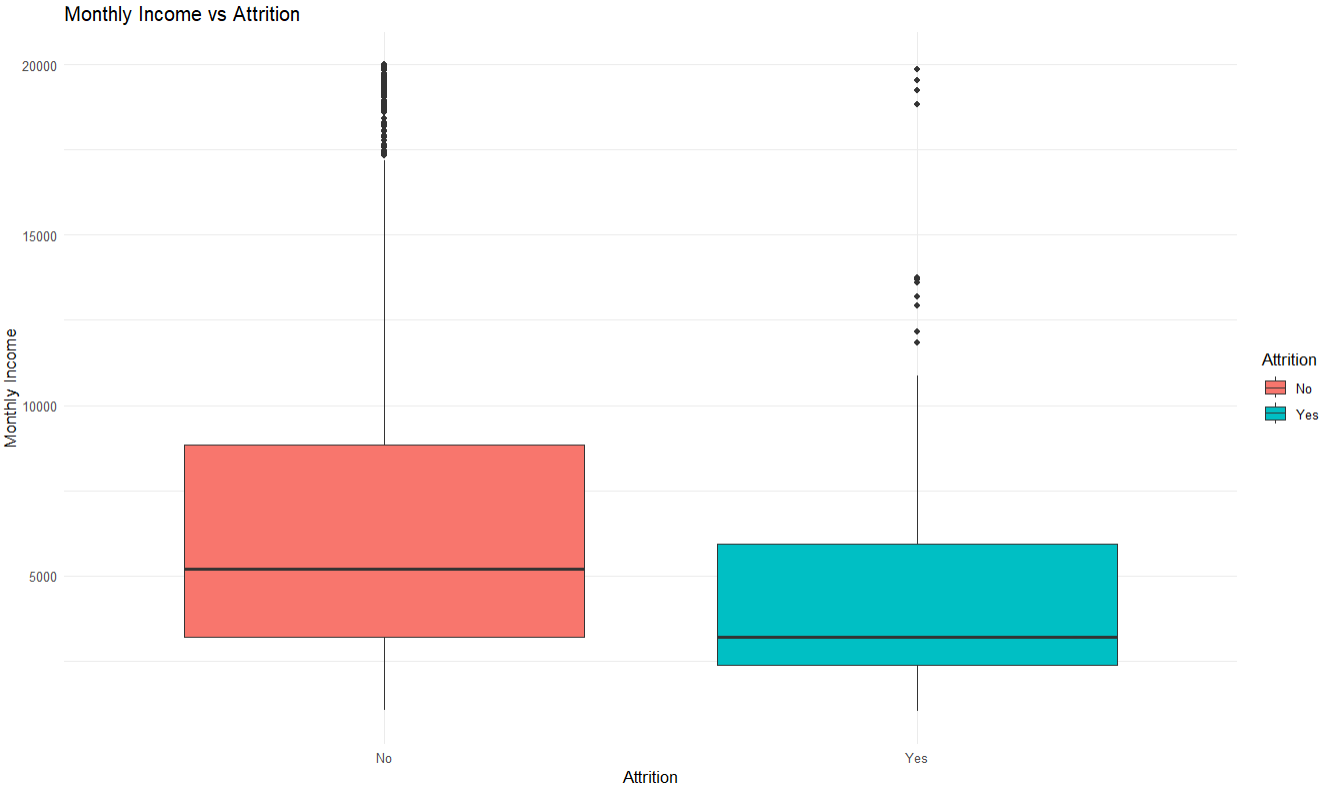
**Distribution of Employee Salaries by Department**

In the pie chart illustrating salary distribution by department, it is observed that the majority of employees are in the Research and Development department, followed by the Sales department, with the Human Resources department having the fewest employees. This distribution indicates that the company primarily focuses on R&D and Sales, requiring more human resources in these areas. The relatively smaller share of the Human Resources department may suggest that it operates with fewer personnel or that the company allocates fewer resources to this area.

****

**Monthly Income and Attrition Relationship**

The boxplot compares the monthly income distribution between employees who stayed (Attrition = No) and those who left (Attrition = Yes). The data suggests that employees who did not leave (red box) have a higher median monthly income than those who left (blue box). Additionally, the income range is broader for employees who stayed, with more outliers at the higher end. This trend may indicate that employees with higher salaries are less likely to leave their jobs.

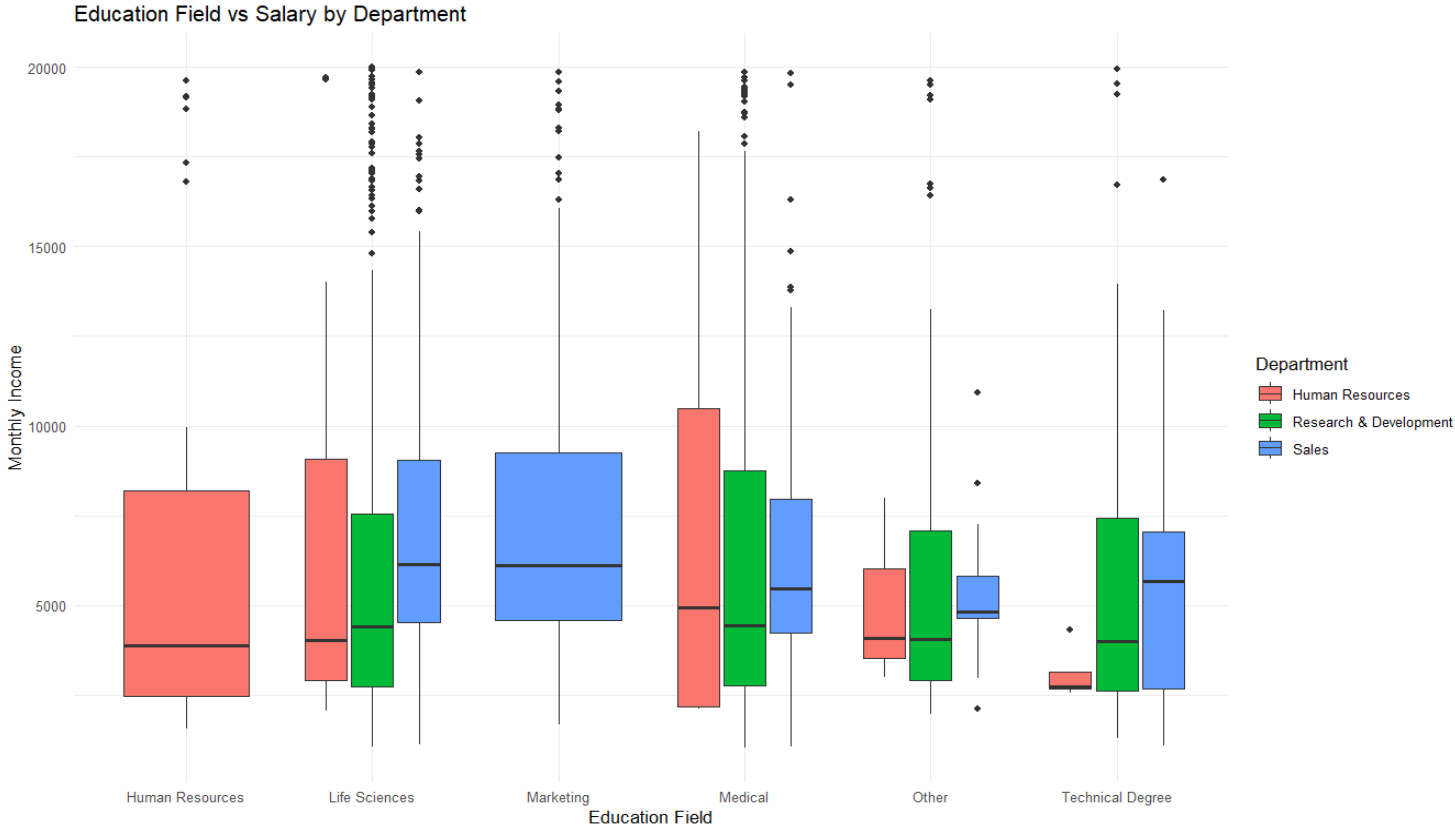
****

**Comprehensive Analysis of Education Field, Department, and Salary Distribution Analysis**

The analysis reveals significant differences in employees' salaries based on their **education field and department**. In general, **Sales and R&D departments** offer higher salaries, while the **Human Resources department** has lower salary levels.

**Impact of Education Field:**

* Employees with **Technical and Medical degrees** tend to earn higher salaries, particularly in **R&D and Sales departments**.
* Those from **Marketing and Life Sciences backgrounds** achieve the highest salary levels, especially in the **Sales department**.
* Salaries in the **Human Resources department** are generally lower across all education fields.

****

**Overall Assessment:**

* **R&D department** has the broadest salary distribution, with high salary levels for employees in technical fields.
* **Sales department** offers the highest salaries in certain education fields (Marketing, Life Sciences).
* **Human Resources department** generally has lower salary ranges, making it a less attractive option for employees with technical degrees.

In this study, a logistic regression model was developed to predict employee attrition, and the model’s performance was evaluated. First, the dataset was split into training and testing sets using the createDataPartition() function, with 70% of the data allocated for training and 30% for testing. Since the data was imbalanced, a careful approach was taken during the partitioning process. Next, a logistic regression model was initially created using the glm() function, including all variables. To improve the model, the stepAIC() method was applied, which eliminated insignificant variables and resulted in a more optimized model.

The refined model (model2) was then used to make predictions on the test data. The predicted values were initially obtained as probabilities. To enhance the model’s predictive performance, a ROC curve was plotted, and the best threshold value was determined using the Youden index method. The predicted probabilities were then classified as either "Yes" (Attrition) or "No" (No Attrition) based on the chosen threshold. Finally, a Confusion Matrix was generated to evaluate the model’s performance, and key metrics such as accuracy, sensitivity, and specificity were calculated. At the end of this process, the attrition probability was predicted as accurately as possible, and the effectiveness of the model was assessed.

**CONCLUSION**

The model accurately predicts employees who stay (**95.6% NPV**) but struggles to identify those who leave (**46.72% PPV**). Although sensitivity (**80.28%**) is high, the model still misclassifies many attrition cases. The overall accuracy (**82.05%**) may be misleading due to class imbalance. A more balanced dataset or alternative modeling approaches could improve attrition detection.

The Kappa statistic, which measures the agreement between predicted and actual classes, is 0.4858. This is a moderate value, indicating a moderate level of agreement beyond chance. The McNemar's test p-value of 1.85e-08 indicates a significant difference between the two classes, which suggests that the model's errors in classification are not equally distributed between the two classes.

For job attrition prediction, the model is effective at detecting employees who stay (negative class), but may struggle to predict those who will leave (positive class). The moderate Kappa and positive predictive value suggest that further improvements could be made, especially in predicting attrition. Given the lower prevalence of attrition (16.14%), the model could benefit from better handling of the rare positive class. However, with high sensitivity and specificity, the model can be more effective if it focuses on improving false positives.

The AUC value of 0.8746 indicates a strong model performance, as it is closer to 1, which suggests that the model is effective in distinguishing between the positive and negative classes. The ROC curve's shape further confirms this, as it typically shows good separation between classes when the AUC is high.