Name : Shaktikalp Mohanty

**Employee Attrition Analysis and Prediction**

This project aims to provide insights into the factors influencing employee attrition and predict which employees are likely to leave the company.

# Problem Statement:

Acme Corporation, a leading tech company, is facing a significant challenge with employee turnover. The HR department is concerned about the increasing rate of attrition, as it negatively impacts team dynamics, project continuity, and overall company morale. To address this issue, Acme Corporation wants to leverage data analytics and machine learning to understand the factors influencing employee turnover and predict

which employees are likely to leave in the near future.

# Dataset:

The dataset typically includes several features:

* 1. **Employee ID:** A unique identifier for each employee.
  2. **Age:** The age of the employee.
  3. **Attrition:** A binary variable indicating whether the employee has left the company (1) or is still employed (0).
  4. **Business Travel:** The frequency and nature of business-related travel (e.g., "Travel\_Rarely," "Travel\_Frequently," "Non-Travel").
  5. **Department:** The department to which the employee belongs (e.g., "Sales," "Research & Development," "Human Resources").
  6. **Distance From Home:** The distance of the employee's residence from the workplace.
  7. **Education:** The employee's level of education (e.g., "1: 'Below College'," "2: 'College'," "3: 'Bachelor'," "4: 'Master'," "5: 'Doctor').
  8. **Education Field:** The field in which the employee's education lies (e.g., "Life Sciences," "Medical," "Marketing").
  9. **Environment Satisfaction:** The level of satisfaction with the work environment on a scale.
  10. **Gender:** The gender of the employee.
  11. **Job Involvement:** The degree to which the employee is involved in their job.
  12. **Job Level:** The level or rank of the employee's position.
  13. **Job Role:** The specific role or title of the employee's job.
  14. **Job Satisfaction:** The level of satisfaction with the job on a scale.
  15. **Marital Status:** The marital status of the employee.
  16. **Monthly Income:** The monthly salary of the employee.
  17. **Num Companies Worked:** The number of companies the employee has worked for.
  18. **Over Time:** Whether the employee works overtime or not.
  19. **Performance Rating:** The performance rating of the employee.
  20. **Relationship Satisfaction:** The level of satisfaction with relationships at the workplace.
  21. **Stock Option Level:** The level of stock options provided to the employee.
  22. **Total Working Years:** The total number of years the employee has been working.
  23. **Training Times Last Year:** The number of training sessions the employee attended last year.
  24. **Work-Life Balance:** The balance between work and personal life.
  25. **Years At Company:** The number of years the employee has been with the current company.
  26. **Years In Current Role:** The number of years the employee has been in their current role.
  27. **Years Since Last Promotion:** The number of years since the last time the employee was promoted.
  28. **Years With Current Manager:** The number of years the employee has been working under the current manager.

# First Tasks: Data Preprocessing and Cleaning:

## DATA EXPLORATION:

* The code begins by importing the necessary libraries, including pandas for data manipulation and scikit-learn for machine learning models.
* The dataset WA\_Fn-UseC\_-HR-Employee-Attrition.csv is loaded into a pandas DataFrame called data.
* The first few rows of the dataset are displayed to get a glimpse of the data.
* Summary statistics of numerical and categorical columns are calculated and displayed using the describe() method.
* The code checks for missing values in the dataset, and it is confirmed that there are no missing values.

### Basic Information:

* + The dataset contains 1470 entries and 35 columns.
  + No missing values in the dataset.
  + Columns include a mix of numerical and categorical data.

### Summary Statistics:

* + Key statistics for numerical columns (mean, std, min, max, etc.) show typical ranges for employee- related data.

### Missing Values:

* + There are no missing values in any columns.

### Unique Value Counts:

* + Certain columns have high cardinality (e.g., EmployeeNumber), while others have limited unique values (e.g., Gender, Attrition).

## DATA CLEANING

Since there are no missing values, the next step is to check for inconsistent or outlier values and handle them accordingly. However, it seems the data is relatively clean.

Some columns have only one unique value across all rows, meaning they do not provide any useful information for the model. They are constant and do not contribute to the variance in the data. So, we dropped that column.

x = df.drop(['EmployeeCount','Over18','StandardHours'], axis=1)

## DATA ENCODING

*  The categorical columns in the dataset are encoded using different techniques to convert them into a format suitable for machine learning algorithms.
* The Attrition column is encoded using ordinal encoding, where 'Yes' is mapped to 1 and 'No' is mapped to 0.
* The BusinessTravel, Department, EducationField, Gender, JobRole, Over18, and MaritalStatus columns are encoded using label encoding from scikit-learn's LabelEncoder.
* The OverTime column is encoded using binary encoding, where 'No' is mapped to 0 and 'Yes' is mapped to 1.
* After encoding, the first few rows of the encoded dataset are displayed.

We need to convert categorical variables into numerical form.

Columns identified for encoding are: Attrition, BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, Over18, OverTime.

**Label Encoding** : Converts binary categorical features into numerical form.

**One-Hot Encoding** : Converts categorical features with more than two unique values into multiple binary columns.

## DATA LABELLING

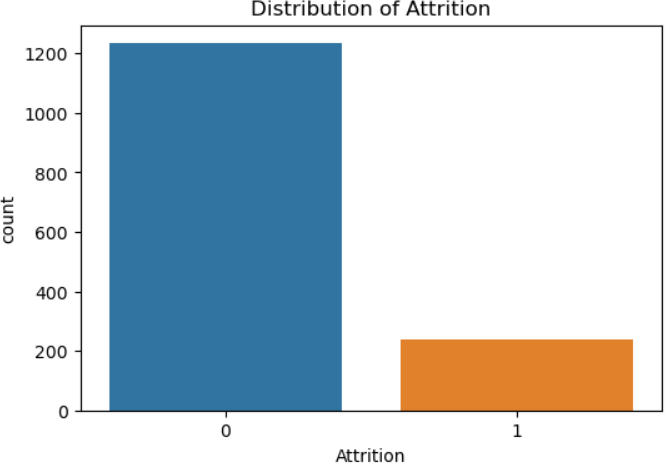
Ensuring that the target variable (Attrition) is in the correct format for modeling.

target\_variable = 'Attrition' label\_encoder = LabelEncoder()

df[target\_variable] = label\_encoder.fit\_transform(df[target\_variable])

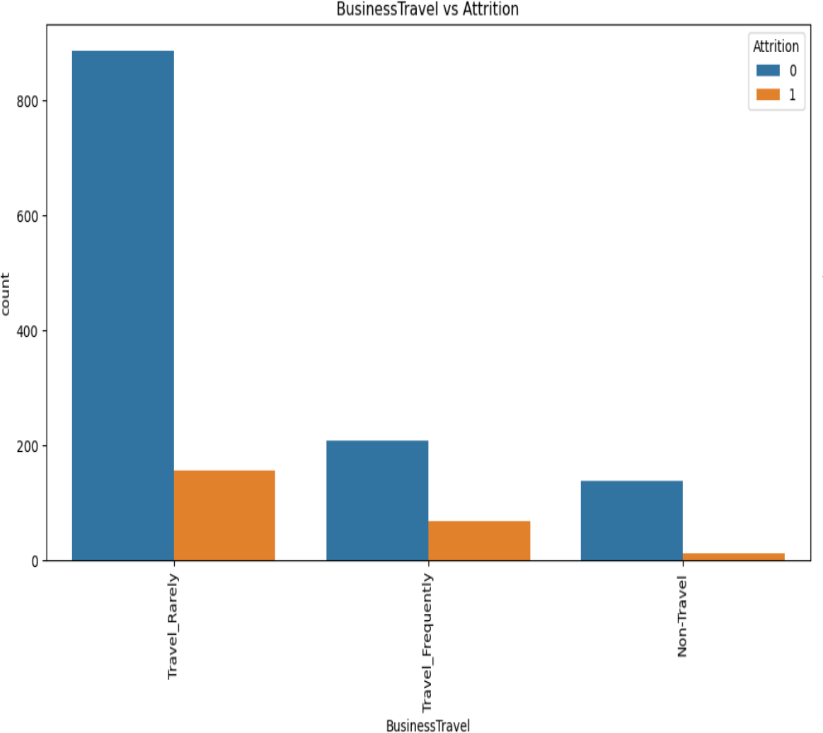
df.head()

## INTERPRETATION OF VISUALIZATION

The first bar chart shows the distribution of Employee Attrition.

* + **0 (No Attrition)**: There are significantly more employees who have not left the organization. This bar is much higher, indicating that a majority of employees are retained.
  + **1 (Attrition)**: The number of employees who have left the organization is relatively small compared to those who stayed.

The **next visualization** is Attrition rates among different Business travel categories.



### Travel\_Rarely:

* Majority of the employees who travel rarely stay in the company.
* However, a noticeable number of them also leave, but the retention rate is still higher.

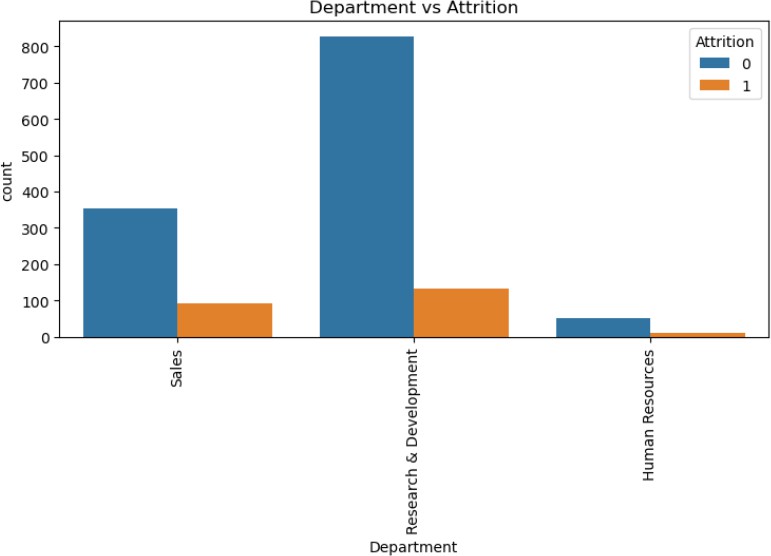
### Travel\_Frequently:

* More employees in this category leave the company compared to those who stay.
* Attrition is higher for frequent travelers.

### Non-Travel:

* The number of non-traveling employees who stay is higher, though fewer employees fall into this category overall.
* Attrition is relatively low.

The **next visualization** compares the Attrition rates among different Departments.



### Sales:

* Significant number of employees stay, but there is also a considerable number who leave.

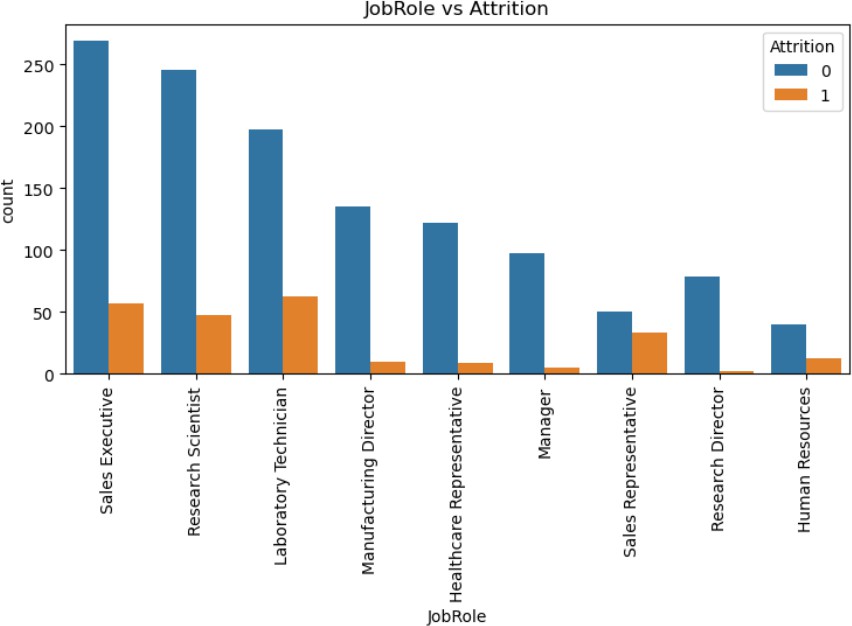
### Research & Development:

* The majority of employees stay, with attrition being relatively lower.

### Human Resources:

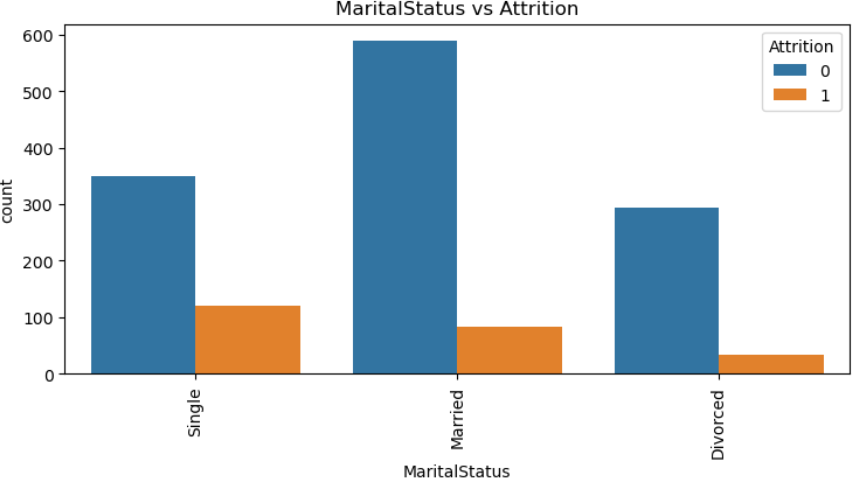
* Fewer employees in this department overall.
* Both retention and attrition numbers are low, but attrition appears to be slightly higher in proportion to the department size.

The **next visualization** compares the Attrition rates among different Job Roles.



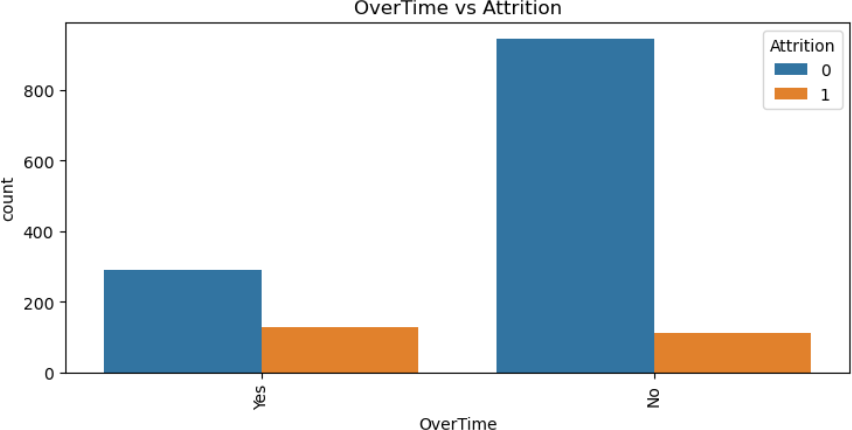
* S**ales Executive**: This role has the highest number of employees, with a noticeable number of employees experiencing attrition.
* **Research Scientist** and **Laboratory Technician**: These roles also have significant employee counts, with a moderate level of attrition.
* **Healthcare Representative** and **Sales Representative**: These roles show lower employee counts but higher proportions of attrition.
* **Manager** and **Manufacturing Director**: Lower levels of attrition are observed in these roles.
* **Research Director** and **Human Resources**: These roles have the lowest attrition rates.

The **next visualization** compares the Attrition rates among Marital Status.



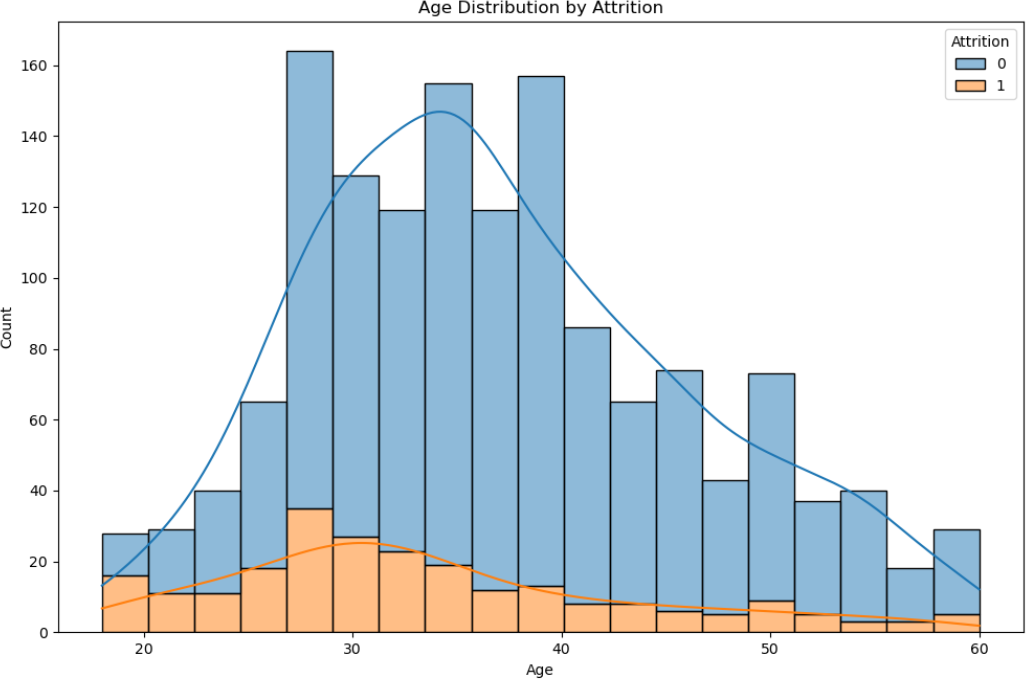
* **Single**: Single employees show a higher level of attrition compared to their married and divorced counterparts.
* **Married**: Married employees have the highest count but the lowest proportion of attrition.
* **Divorced**: Divorced employees show a lower count with moderate attrition rates.

The **next visualization** compares the Attrition rates vs OverTime.



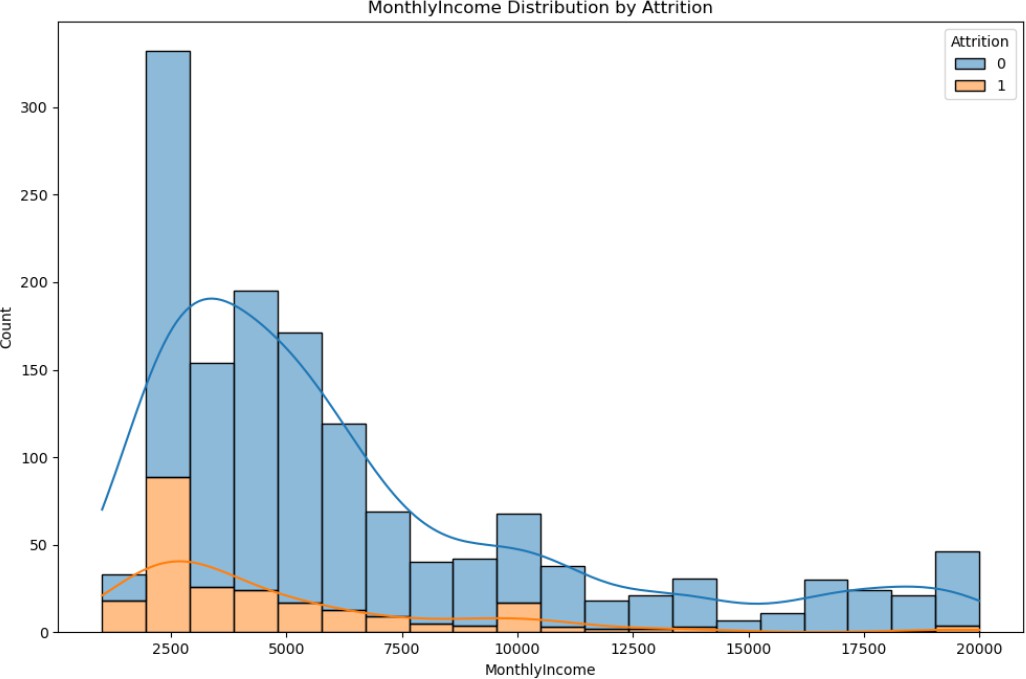
* **Yes (Overtime)**: Employees working overtime show a significant level of attrition.
* **No (Overtime)**: Employees not working overtime have a higher count overall, with a lower proportion of attrition compared to those working overtime.

The **next visualization** compares the Attrition rates among different Age.



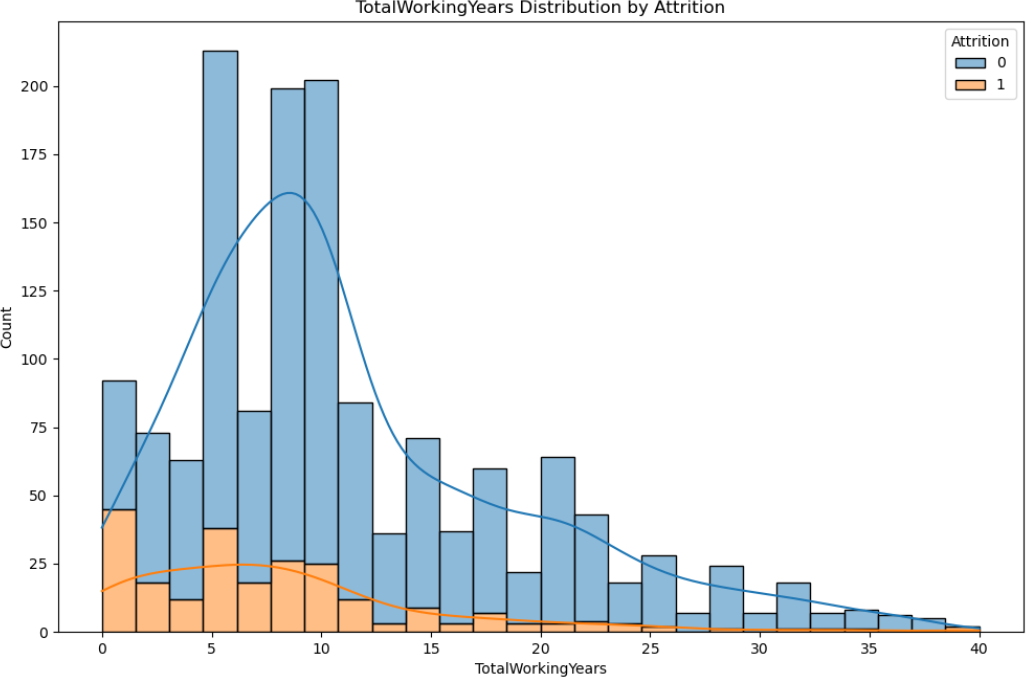
* The majority of employees are in the 30-40 age range.
* Attrition seems to be higher in the younger age group (20-30) and decreases with age, but then shows a slight increase around the 40-50 range.

The **next visualization** compares the attrition rates among Monthly Income.



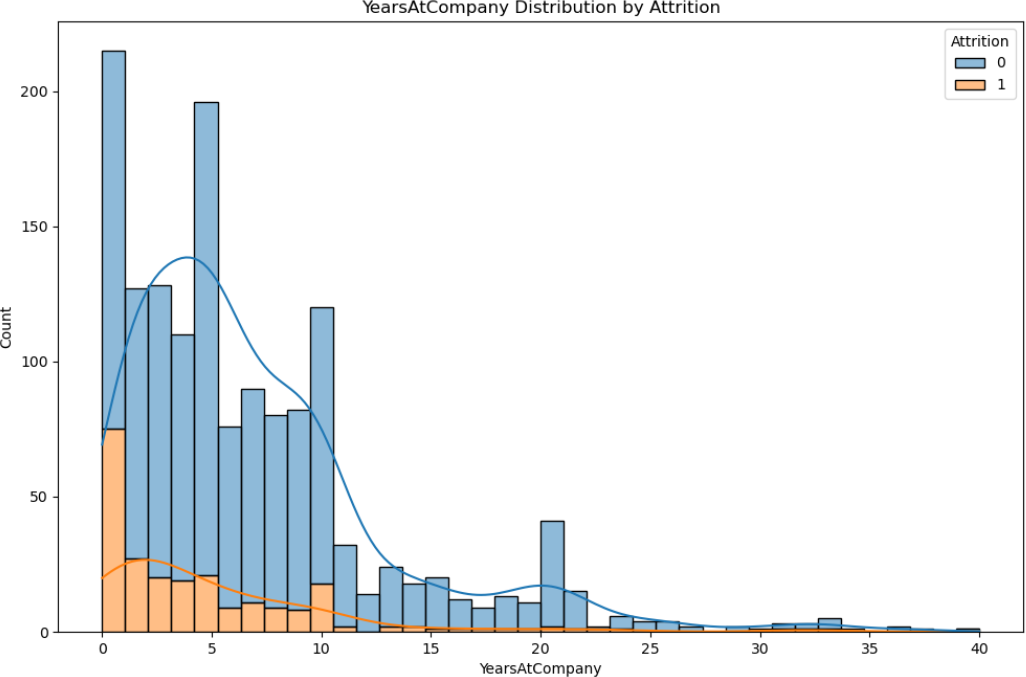
* Most employees have a lower monthly income, with the highest concentration between 0 and 5,000 units.
* The number of employees decreases as the monthly income increases.
* Attrition (orange) is more noticeable among employees with lower monthly income (0 to 5,000 units).
* As the monthly income increases, the rate of attrition decreases significantly.
* Employees with higher incomes (above 10,000 units) have minimal attrition.

The **next visualization** is TotalWorkingYears distribution by Attrition.



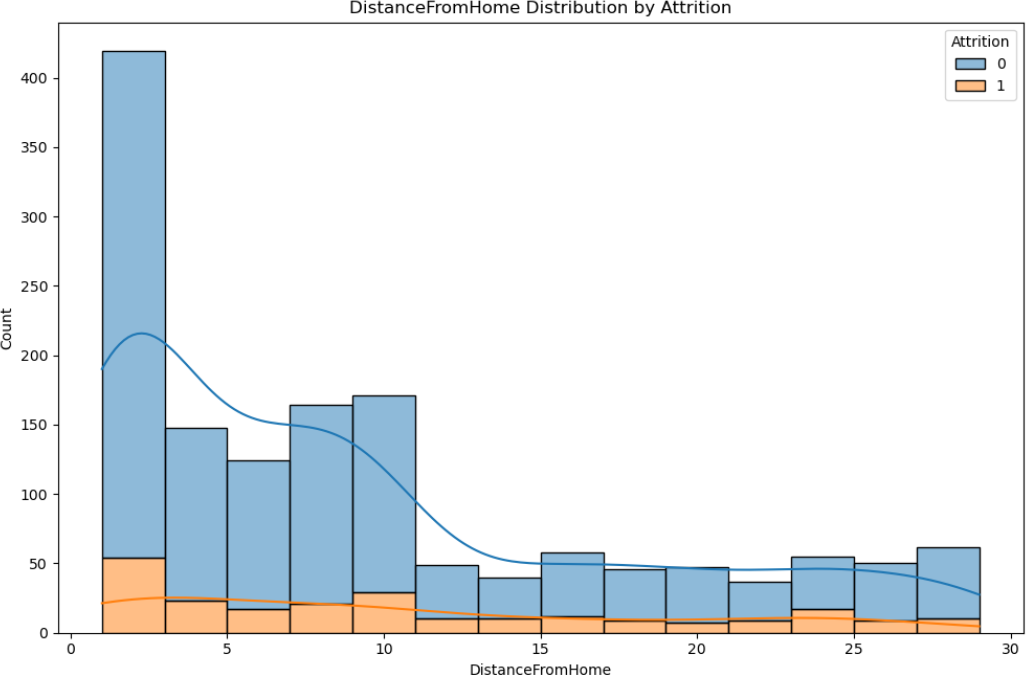
* A higher concentration of employees with lower total working years, particularly in the range of 0 to 15 years.
* The peak for non-attrition (blue) is around 8-10 years.
* Attrition (orange) is relatively high among those with 0-5 years of total working experience and then decreases as the total working years increase.
* Very few employees with total working years beyond 20 years have left, indicating a potential retention of experienced employees.

The **next visualization** is YearsAtCompany distribution by Attrition .



* A significant number of employees have 0-5 years at the company, with a noticeable peak in the 0-2 year range.
* Attrition (orange) is highest among employees who have spent 0-2 years at the company, suggesting new hires are more likely to leave.
* After the 5-year mark, attrition decreases significantly and remains low for employees with longer tenure.
* There is a secondary peak around the 20-year mark for non-attrition (blue), indicating long-term employees tend to stay.

The **next visualization** is DistanceFromHome distribution by Attrition.



* Most employees live within a short distance from home (0 to 5 units).
* The number of employees decreases as the distance from home increases.
* Attrition (orange) appears to be relatively low across all distances.
* There is a slight increase in attrition at shorter distances (0 to 5 units), which decreases progressively as the distance increases.

**Conclusion**

The analysis of employee attrition reveals:

* **Overall Retention:** Most employees stay with the company.
* **Income:** Higher incomes correlate with lower attrition.
* **Business Travel:** Frequent travellers have higher attrition; non-travellers are more likely to stay.
* **Departmental Impact:** Research & Development has the highest retention, followed by Sales, then Human Resources.
* **Job Role:** High attrition in Sales Executives, Research Scientists, and Laboratory Technicians.
* **Marital Status:** Single employees are more likely to leave than married or divorced ones.
* **Overtime:** Employees working overtime have higher attrition rates.
* **Age:** Younger employees have higher attrition.
* **Monthly Income:** Lower incomes are associated with higher attrition.
* **Early Years:** Attrition is highest among employees in their first 0-5 years at the company.
* **Experienced Employees:** Lower attrition among long-tenured staff.
* **Distance from Home:** Slightly higher attrition for employees living closer to work, though overall attrition is low across all distances.

## MODEL

 The code splits the dataset into training and testing sets using scikit-learn's train\_test\_split function.

 A logistic regression model is created using LogisticRegression from scikit-learn.

 The model is fitted on the training data using the fit() method.

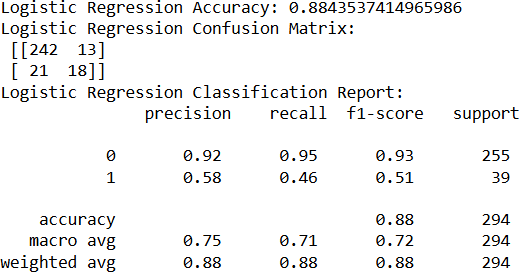
 Predictions are made on the test data using the predict() method.

 The actual and predicted values are combined into a pandas DataFrame called result.

 The confusion matrix is calculated and displayed using scikit-learn's confusion\_matrix function.

 The classification report, which includes precision, recall, F1-score, and accuracy, is generated and displayed using scikit-learn's classification\_report function.

* Model : **LOGISTIC REGRESSION**



The Logistic Regression model performed well, achieving an accuracy of 88.4%.

### Recall (Sensitivity):

* The recall for class 1 (attrition) is 0.46, meaning the model only catches 46% of employees who actually leave the company. In other words, it misses more than half of the employees who actually quit.

### Precision:

* The precision for class 1 is 0.58, indicating that when the model predicts attrition, it's correct about 58% of the time. So, out of all the times the model flags attrition, roughly 58% of them are actual cases of employees leaving.

However, the model shows lower performance in predicting the minority class (attrition). The recall for class 1 (attrition) is 0.46, indicating that the model correctly identifies 46% of actual attrition cases. The precision for class 1 is 0.58, meaning that when the model predicts attrition, it is correct 58% of the time.

**Best Model**: Logistic Regression

* **Reason**: Logistic Regression provides a balanced trade-off between precision and recall for both classes. Logistic Regression still offers the most balanced performance, especially important for the minority class (class 1).