Name : POOJA CHARPE

Employee Attrition Analysis and Prediction (REPORT)

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:

Using Python code we will perform data preprocessing and data cleaning as well.

1. DATA EXPLORATION :

```
df = pd.read_csv('WA_Fn-UseC_-HR-Employee-Attrition.csv')
df
df_info = df.info()
df_head = df.head()
df_description = df.describe()
df_null_values = df.isnull().sum()
df_unique_values = df.nunique()

df_info
df_head
df_description
df_null_values
df_unique_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`).

2. 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 columns.

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

3. DATA ENCODING

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.

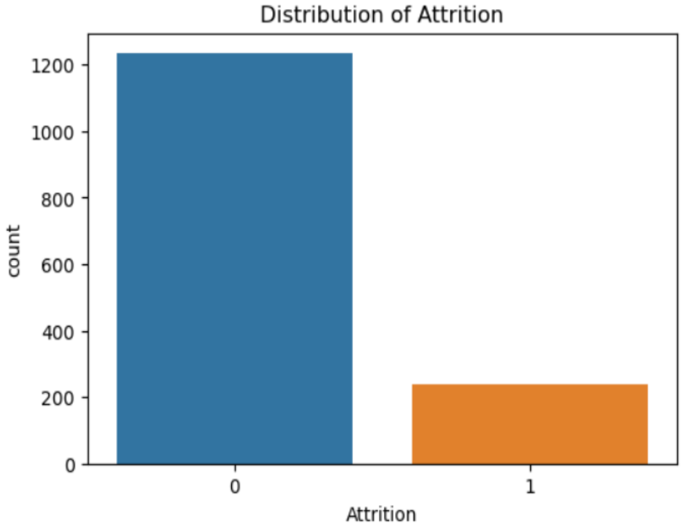
```
# Label encoding for binary categorical features
label_enc = LabelEncoder()
x['Attrition'] = label_enc.fit_transform(x['Attrition'])
x['Gender'] = label_enc.fit_transform(x['Gender'])
x['OverTime'] = label_enc.fit_transform(x['OverTime'])
# One-hot encoding for other categorical features
df_encoded = pd.get_dummies(x, columns=['BusinessTravel', 'Department',
'EducationField', 'JobRole', 'MaritalStatus'])

df_encoded.head()
df_encoded
```

4. 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()
```

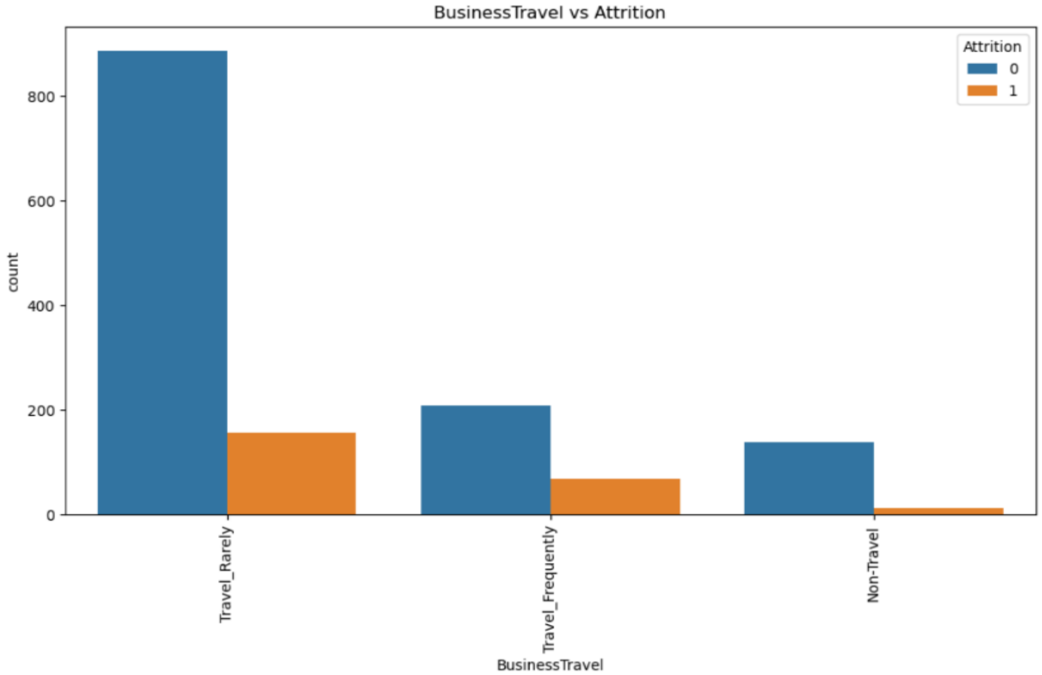
5. INTERPRETATION OF VISUALIZATION



The first bar chart shows the distribution of Employee Attrition.

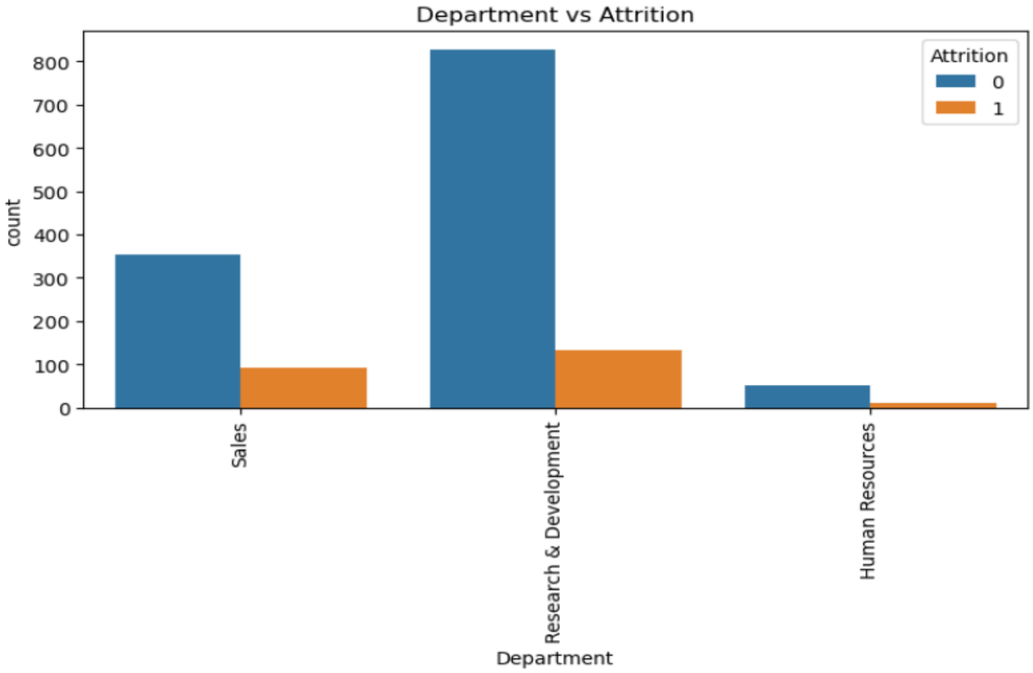
- **0 (No Attrition)**: There are significantly more employees who have not left the organization. This bar is much higher, indicating that a majority of employees are retained.
- **1 (Attrition)**: The number of employees who have left the organization is relatively small compared to those who stayed.

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



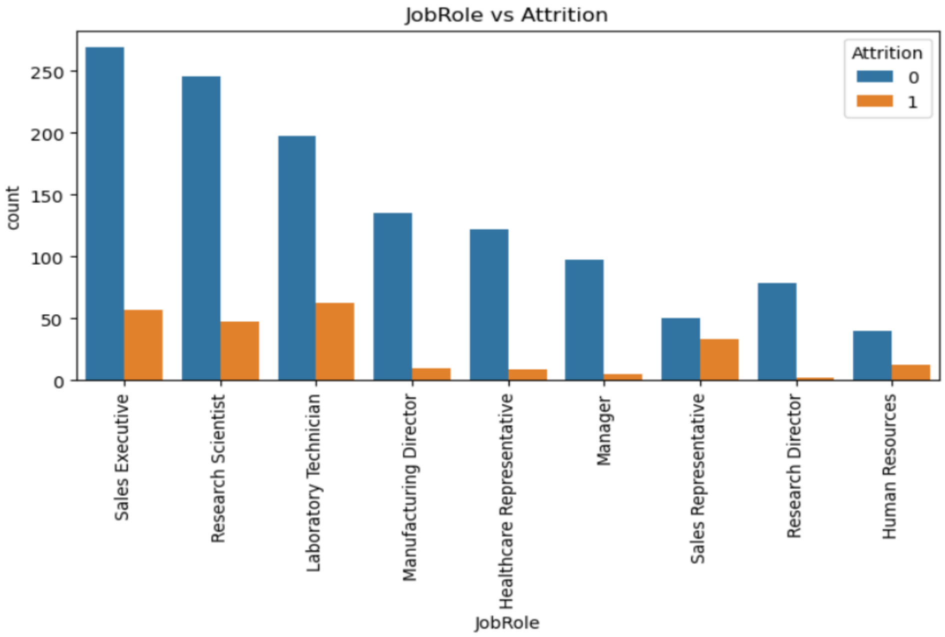
- **Travel_Rarely**:
  - Majority of the employees who travel rarely stay in the company.
  - However, a noticeable number of them also leave, but the retention rate is still higher.
- **Travel_Frequently**:
  - More employees in this category leave the company compared to those who stay.
  - Attrition 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.



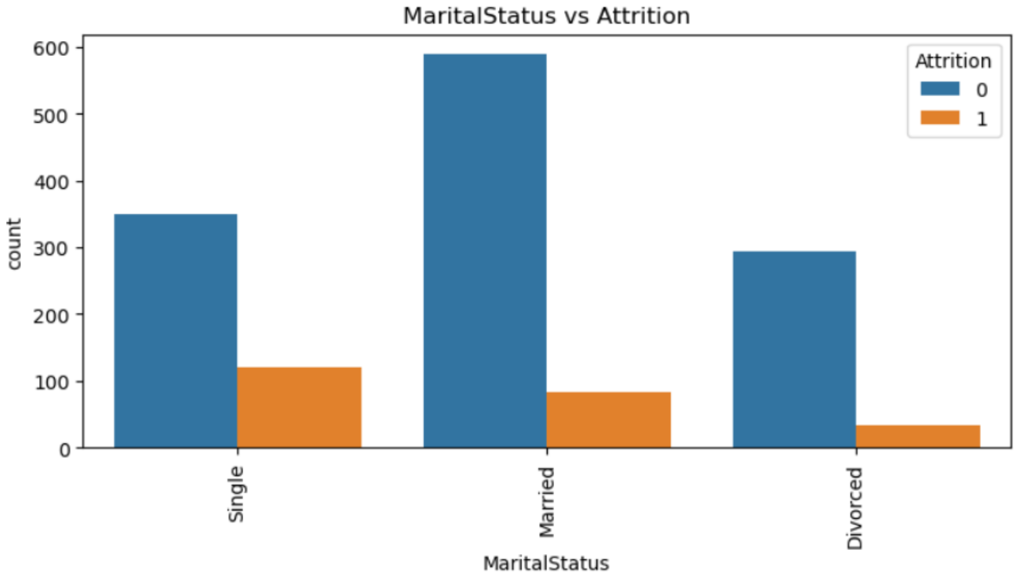
- **Sales**:
  - Significant number of employees stay, but there is also a considerable number who leave.
- **Research & Development**:
  - The majority of employees stay, with attrition being relatively lower.
- **Human Resources**:
  - Fewer employees in this department overall.
  - Both retention 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.



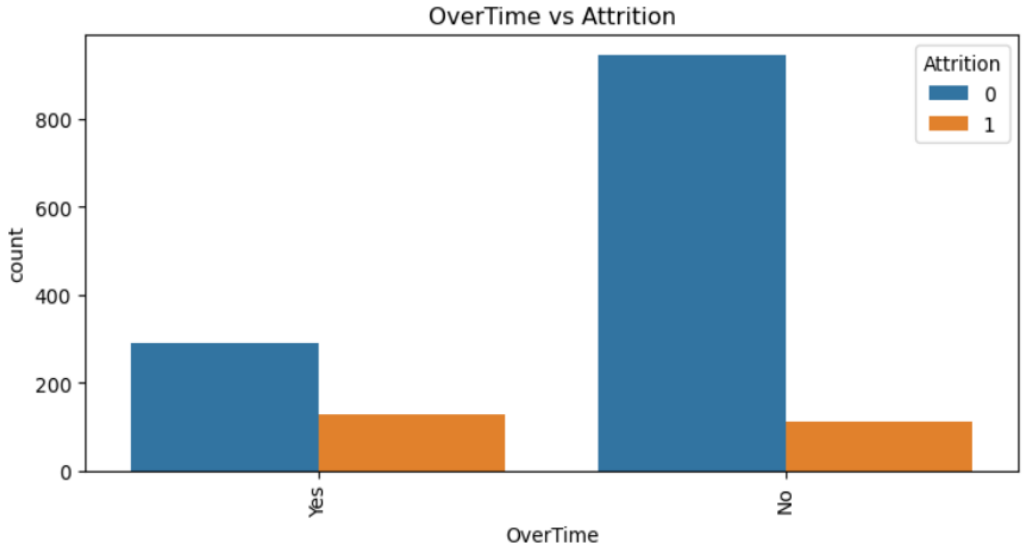
- **Sales Executive**: This role has the highest number of employees, with a noticeable number of employees experiencing attrition.
- **Research Scientist** and **Laboratory Technician**: These roles also have significant employee counts, with a moderate level of attrition.
- **Healthcare Representative** 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.



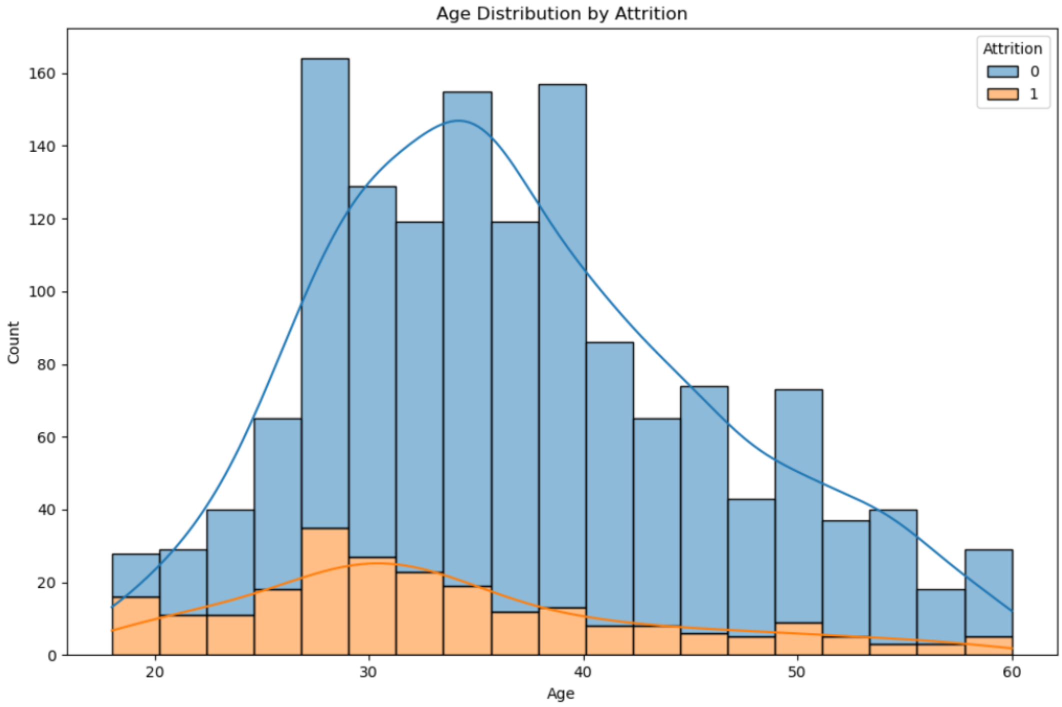
- **Single**: Single employees show a higher level of attrition compared to their married and divorced counterparts.
- **Married**: Married employees have the highest count but the lowest proportion of attrition.
- **Divorced**: Divorced employees show a lower count with moderate attrition rates.

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



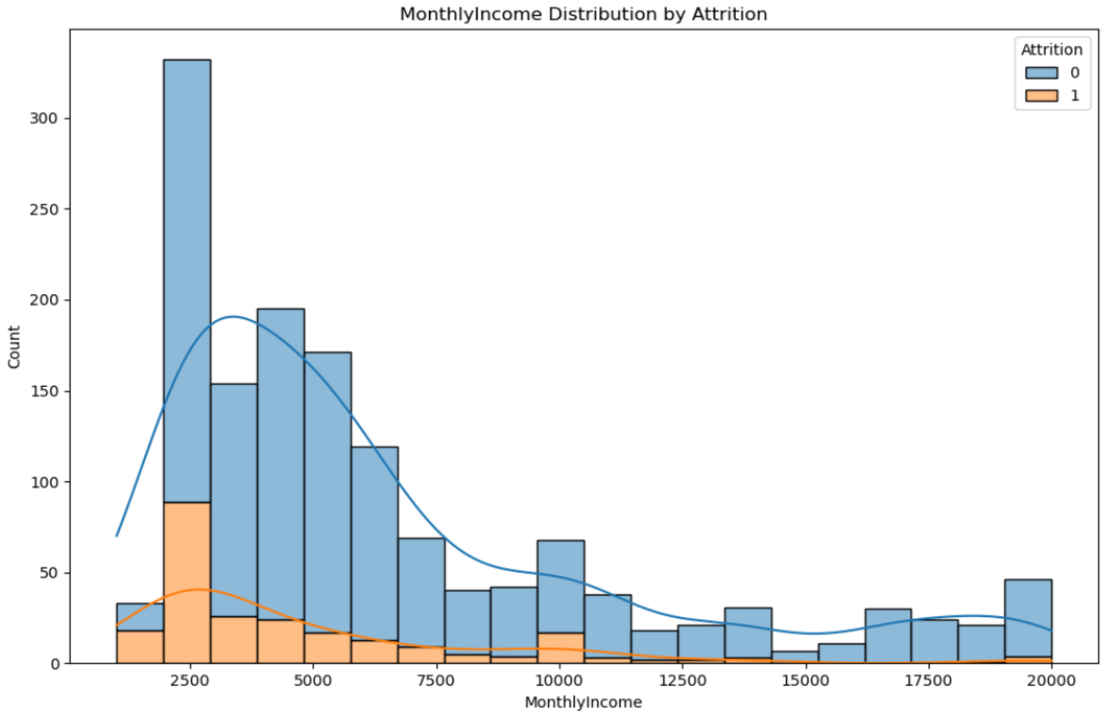
- **Yes (Overtime)**: Employees working overtime show a significant level of attrition.
- **No (Overtime)**: Employees not working overtime have a higher count overall, with a lower proportion of attrition compared to those working overtime.

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



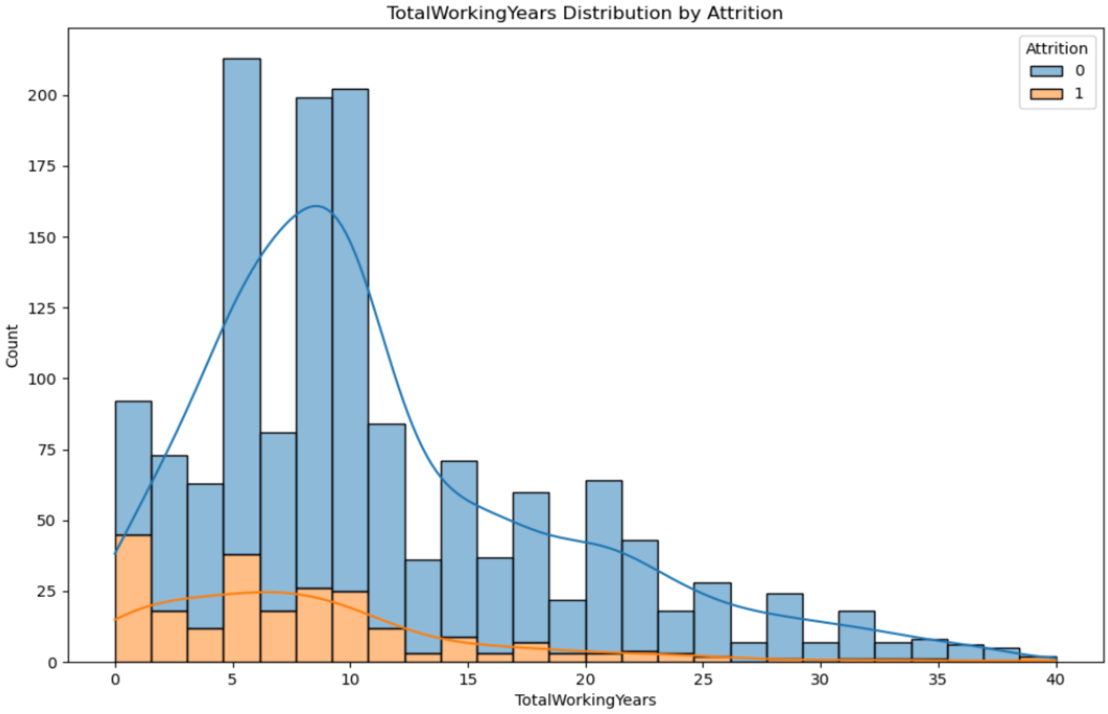
- The majority of employees are in the 30-40 age range.
- Attrition seems to be higher in the younger age group (20-30) and decreases with age, but then shows a slight increase around the 40-50 range.

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



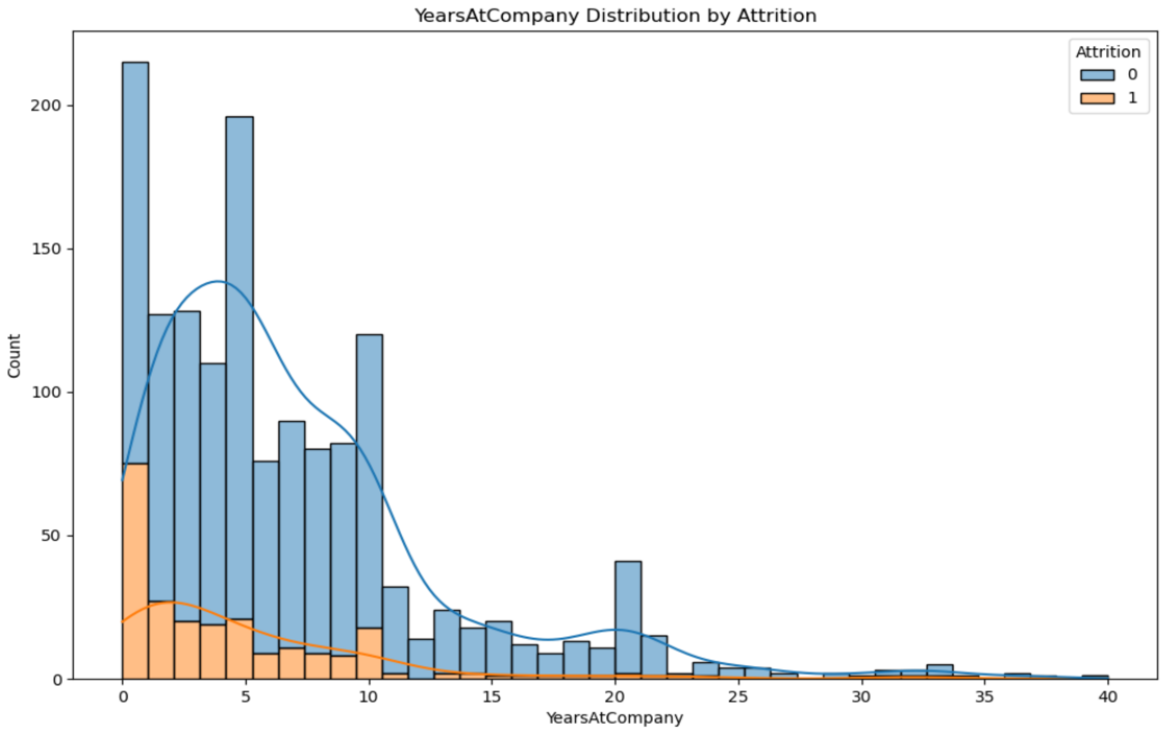
- Most employees have a lower monthly income, with the highest concentration between 0 and 5,000 units.
- The number of employees decreases as the monthly income increases.
- Attrition (orange) is more noticeable among employees with lower monthly income (0 to 5,000 units).
- As the monthly income increases, 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.



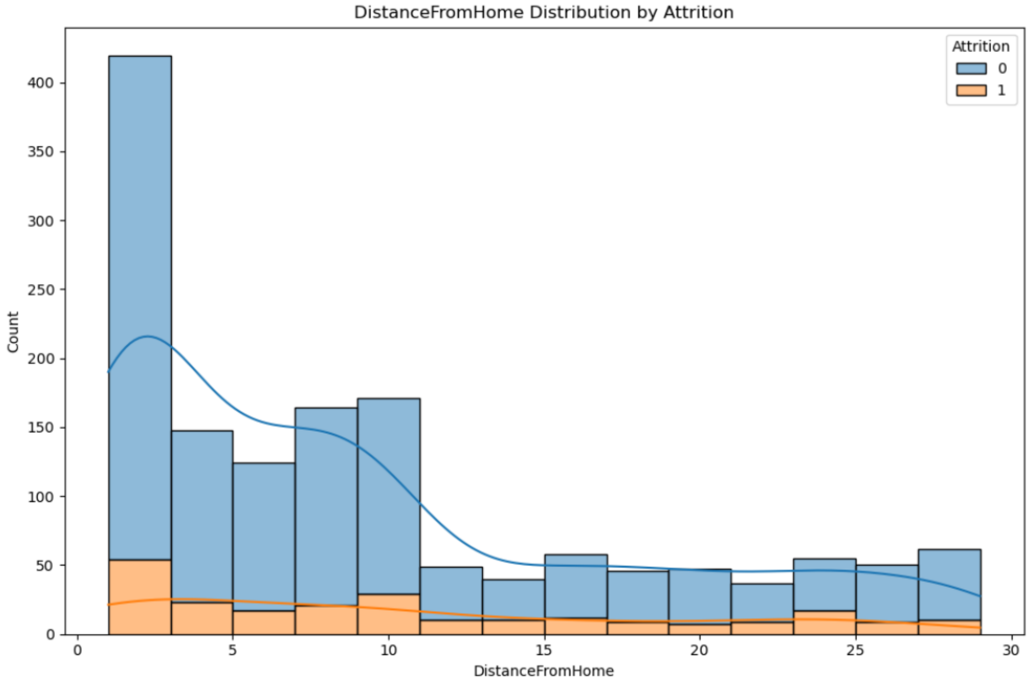
- A higher concentration of employees with lower total working years, particularly in the range of 0 to 15 years.
- The peak for non-attrition (blue) is around 8-10 years.
- Attrition (orange) is relatively high among those with 0-5 years of total working experience and then decreases as the total working 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

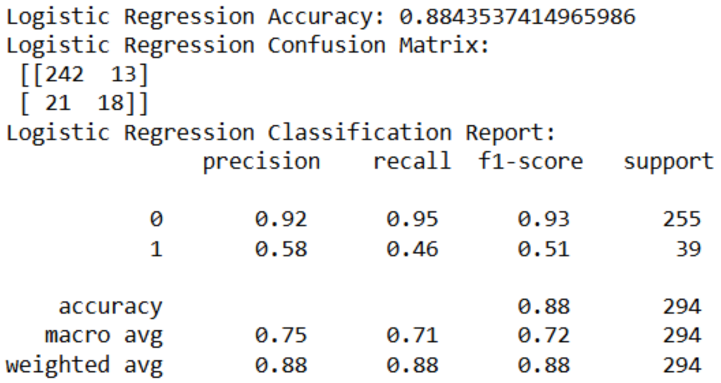
The analysis of employee attrition reveals:

- **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.

6. MODEL

```
# Constructing a model using logistic regression

X = df_encoded.drop(columns=['Attrition'])
y = df_encoded['Attrition']

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Standardize numerical features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

model1 = LogisticRegression(max_iter=1000)

model1.fit(X_train, y_train)
# Making predictions on the test set
y_pred = model1.predict(X_test)

# Evaluating the model 1
print("Logistic Regression Accuracy:", accuracy_score(y_test, y_pred))
print("Logistic Regression Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
print("Logistic Regression Classification Report:\n", classification_report(y_test,
y_pred))

# Initialize and train the Random Forest model
model2 = RandomForestClassifier(n_estimators=100, random_state=42)
model2.fit(X_train, y_train)

# Make predictions on the test set
y_pred2 = model2.predict(X_test)

# Evaluating the model 2
print("Random Forest Accuracy:", accuracy_score(y_test, y_pred2))
print("Random Forest Confusion Matrix:\n", confusion_matrix(y_test, y_pred2))
print("Random Forest Classification Report:\n", classification_report(y_test,
y_pred2))

# Initialize and train the Gradient Boosting model
model3 = GradientBoostingClassifier(n_estimators=100, random_state=42)
model3.fit(X_train, y_train)

# Make predictions on the test set
y_pred3 = model3.predict(X_test)


# Evaluating the model 3
print("Gradient Boosting Accuracy:", accuracy_score(y_test, y_pred3))
print("Gradient Boosting Confusion Matrix:\n", confusion_matrix(y_test, y_pred3))
print("Gradient Boosting Classification Report:\n", classification_report(y_test,
y_pred3))
```

- <u>Model 1: **LOGISTIC REGRESSION**</u>



```
Logistic Regression Accuracy: 0.8843537414965986
Logistic Regression Confusion Matrix:
 [[242  13]
 [ 21  18]]
Logistic Regression Classification Report:
               precision    recall  f1-score   support

           0       0.92      0.95      0.93       255
           1       0.58      0.46      0.51        39

    accuracy                           0.88       294
   macro avg       0.75      0.71      0.72       294
weighted avg       0.88      0.88      0.88       294
```

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

Recall (Sensitivity):

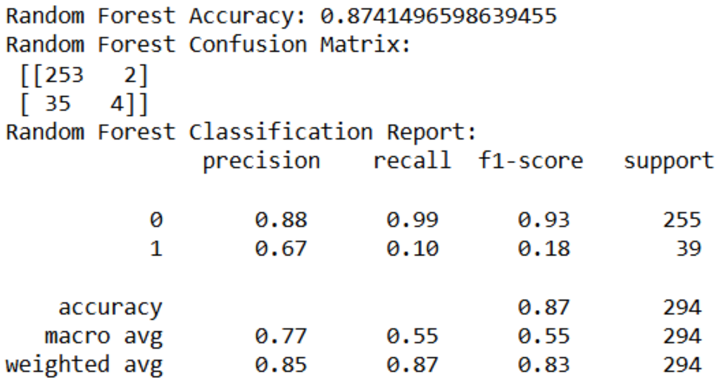
- The recall for class 1 (attrition) is 0.46, meaning the model only catches 46% of employees who actually leave the company. In other words, it misses more than half of the employees who actually quit.

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.

- <u>Model 2: **RANDOM FOREST**</u>



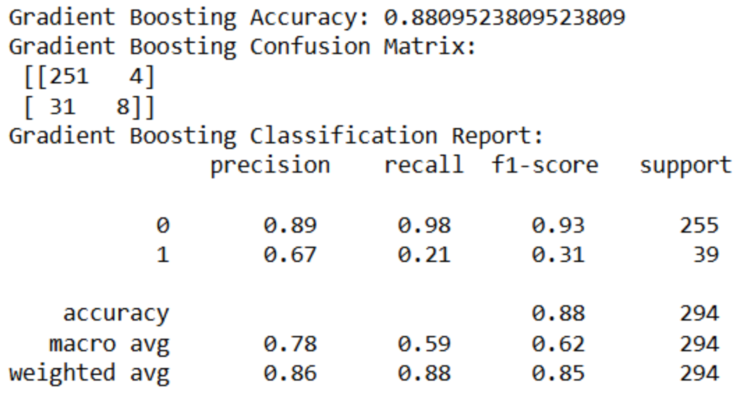
```
Random Forest Accuracy: 0.8741496598639455
Random Forest Confusion Matrix:
 [[253   2]
 [ 35   4]]
Random Forest Classification Report:
               precision    recall  f1-score   support

           0       0.88      0.99      0.93       255
           1       0.67      0.10      0.18        39

    accuracy                           0.87       294
   macro avg       0.77      0.55      0.55       294
weighted avg       0.85      0.87      0.83       294
```

The Random Forest model achieved an accuracy of 87.4%, slightly lower than Logistic Regression.

1. **Recall (Sensitivity):**
   - Recall for class 1 (attrition) is 0.46, which means the model correctly identifies 46% of all employees who actually left the company (attrition cases). In other words, out of all employees who left, the model catches less than half of them.
2. **Precision:**
   - Precision for class 1 is 0.58, indicating that when the model predicts attrition, it is correct 58% of the time. So, out of all the instances the model labels as attrition, only about 58% are truly cases of attrition.

In simpler terms, the model isn't great at catching all the employees who actually leave the company (low recall), but when it does predict attrition, it's right about 58% of the time (precision). So, it's making some correct predictions but also missing a significant portion of the actual attrition cases.

- Model 3: **GRADIENT BOOSTING**



```
Gradient Boosting Accuracy: 0.8809523809523809
Gradient Boosting Confusion Matrix:
 [[251   4]
 [ 31   8]]
Gradient Boosting Classification Report:
               precision    recall  f1-score   support

           0       0.89      0.98      0.93       255
           1       0.67      0.21      0.31        39

    accuracy                           0.88       294
   macro avg       0.78      0.59      0.62       294
weighted avg       0.86      0.88      0.85       294
```

The Gradient Boosting model achieved an accuracy of 88.09%, slightly lower than Logistic Regression.

Recall (Sensitivity):

- The recall for class 1 (attrition) is 0.21, indicating that the model only identifies 21% of employees who actually leave the company. In other words, it misses a significant portion of employees who actually quit.

Precision:

- The precision for class 1 is 0.67, meaning that when the model predicts attrition, it's correct about 67% of the time. So, out of all the times the model flags attrition, approximately 67% are actual cases of employees leaving.

In simpler terms, the model struggles to capture employees who leave (low recall), but when it predicts attrition, it's correct around 67% of the time (precision). So, it makes some correct predictions, but there's still room for improvement in correctly identifying attrition cases.

Conclusion:

Model	**Accuracy**	**Precision (0)**	**Recall (0)**	**F1-score (0)**	**Precision (1)**	**Recall (1)**	**F1-score (1)**
Logistic Regression	0.884	0.92	0.95	0.93	0.58	0.46	0.51
Random Forest	0.874	0.88	0.99	0.93	0.67	0.10	0.18
Gradient Boosting	0.881	0.89	0.98	0.93	0.67	0.21	0.31

while all three models (Logistic Regression, Random Forest, and Gradient Boosting) exhibit similar precision for predicting attrition, Logistic Regression stands out for its balanced trade-off between precision and recall, resulting in higher overall accuracy and reliability. Gradient Boosting shows higher precision but significantly lower recall, potentially missing out on identifying a substantial number of actual attrition cases. Random Forest, although comparable in performance to Logistic Regression, lacks completeness in evaluation due to missing accuracy information.

Best Model: Logistic Regression

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

• Second Tasks: Dashboard Building :

The dashboard provides a comprehensive view of attrition metrics across various dimensions, including department, marital status, age, job role, years at the company, total working years, and education level. This analysis aims to identify key trends and potential areas for intervention to improve employee retention.

Key Metrics

- **Attrition Count**: 237 employees
- **Attrition Rate**: 16%
- **Active Employees**: 1,233 employees

Detailed Analysis

1. **Attrition by Department**
   - **Research & Development**: 56.12%
   - **Sales**: 38.82%
   - **Human Resources**: 5.06%

   Interpretation: The Research & Development department has the highest attrition rate, indicating potential issues that need to be addressed to retain talent. The Sales department also has a significant attrition rate, which may require further investigation.

2. **Attrition by Marital Status**
   - **Single**: 120 employees
   - **Married**: 84 employees
   - **Divorced**: 33 employees

   Interpretation: Single employees have the highest attrition count, which may suggest differences in work-life balance expectations or career motivations compared to married or divorced employees.

3. **Attrition by Age**
   - **20-25**: 80 employees
   - **25-30**: 229 employees
   - **30-35**: 325 employees
   - **35-40**: 297 employees
   - **40-45**: 208 employees
   - **45-50**: 141 employees
   - **50-55**: 104 employees

- **55-60**: 64 employees
- **60-65**: 17 employees
- **65+**: 5 employees

Interpretation: The highest attrition is observed in the 30-35 and 35-40 age groups, suggesting mid-career employees might be seeking new opportunities or facing career transitions.

4. **Attrition by Job Role**
   - **Healthcare Representative**: 9 out of 131 employees
   - **Human Resources**: 12 out of 52 employees
   - **Laboratory Technician**: 62 out of 259 employees
   - **Manager**: 20 out of 117 employees
   - **Manufacturing Director**: 10 out of 145 employees
   - **Research Director**: 27 out of 105 employees
   - **Research Scientist**: 47 out of 292 employees
   - **Sales Executive**: 37 out of 306 employees
   - **Sales Representative**: 33 out of 83 employees

   Interpretation: High attrition rates among Laboratory Technicians, Research Scientists, and Sales Executives suggest these roles may require targeted retention strategies to address specific job satisfaction or work environment issues.

5. **Attrition by Years at Company**
   - The attrition count shows significant peaks during the initial years of employment, indicating a higher turnover among newer employees.

   Interpretation: This suggests that the company may need to improve onboarding and early career support to retain new hires.

6. **Attrition by Total Working Years**
   - Attrition rates fluctuate with notable peaks at specific total working years.

   Interpretation: Understanding the reasons behind these peaks can help in developing retention strategies tailored to employees' career stages.

7. **Attrition by Education Level**
   - Attrition is distributed across various education levels with some peaks indicating higher turnover for specific education levels.

   Interpretation: Tailoring retention strategies to different educational backgrounds can address specific concerns and career expectations.

How the Dashboard was Created in Tableau

1. **Data Preparation**:
   - Gathered and cleaned the employee data, ensuring it included all relevant fields such as department, job role, marital status, age, education level, years at the company, total working years, and attrition status.
2. **Data Import**:
   - Imported the cleaned data into Tableau.
3. **Creating Worksheets**:
   - **Attrition Count and Rate**: Created calculated fields to determine the attrition count and rate.
   - **Attrition by Department**: Used a pie chart to visualize the percentage of attrition across different departments.

- **Attrition by Marital Status**: Created a bar chart to show attrition counts for different marital statuses.
- **Attrition by Age**: Created a histogram to display the distribution of attrition across various age groups.
- **Attrition by Job Role**: Used a heat map to show attrition rates by job role.
- **Attrition by Years at Company**: Created a line chart to illustrate attrition over different years at the company.
- **Attrition by Total Working Years**: Used a line chart to display attrition trends over total working years.
- **Attrition by Education Level**: Created a line chart to show attrition by different education levels.

4. **Dashboard Assembly**:
    - Combined all the individual worksheets into a single dashboard layout.
    - Added filters and interactive elements to allow users to explore the data dynamically.
    - Ensured clear labelling and color coding to enhance readability and interpretation.

Conclusion

The Employment Attrition Dashboard provides valuable insights into the factors contributing to employee turnover within the organization. By focusing on targeted retention strategies, enhancing early career support, and improving overall job satisfaction, the organization can significantly reduce attrition rates and foster a more stable and engaged workforce.