Exploratory Data Analysis (EDA) — HR Analytics Project

1. Libraries Used

- Pandas – data manipulation and cleaning
- NumPy – numerical operations
- Matplotlib & Seaborn – data visualization

2. Data Import and Basic Exploration

- Imported the IBM HR Employee Attrition dataset.
- Viewed the first few rows using df.head() to understand data structure.
- Used df.info() to check datatypes and confirm no missing values.
- Used df.describe() to get summary statistics for numeric columns.
- Checked dataset dimensions using df.shape and listed all column names using df.columns.

Insight: The dataset is well-organized with a balanced mix of numeric and categorical variables, suitable for HR attrition prediction.

3. Data Quality Verification

- Missing Values: Checked with df.isnull().sum() → No null values found.
- Duplicate Values: Checked with df.duplicated().sum() → No duplicate rows found.
- Target Variable (“Attrition”):
  - Verified unique values: 'Yes' and 'No'.
  - Counted class distribution to understand attrition imbalance.

Insight: The dataset is clean, complete, and free from duplication or missing data, ensuring reliability for analysis.

4. Categorical Feature Exploration

Analysed **unique categories and their counts** for all categorical columns:

- BusinessTravel
- Department
- EducationField
- Gender
- JobRole
- MaritalStatus
- OverTime
- Education

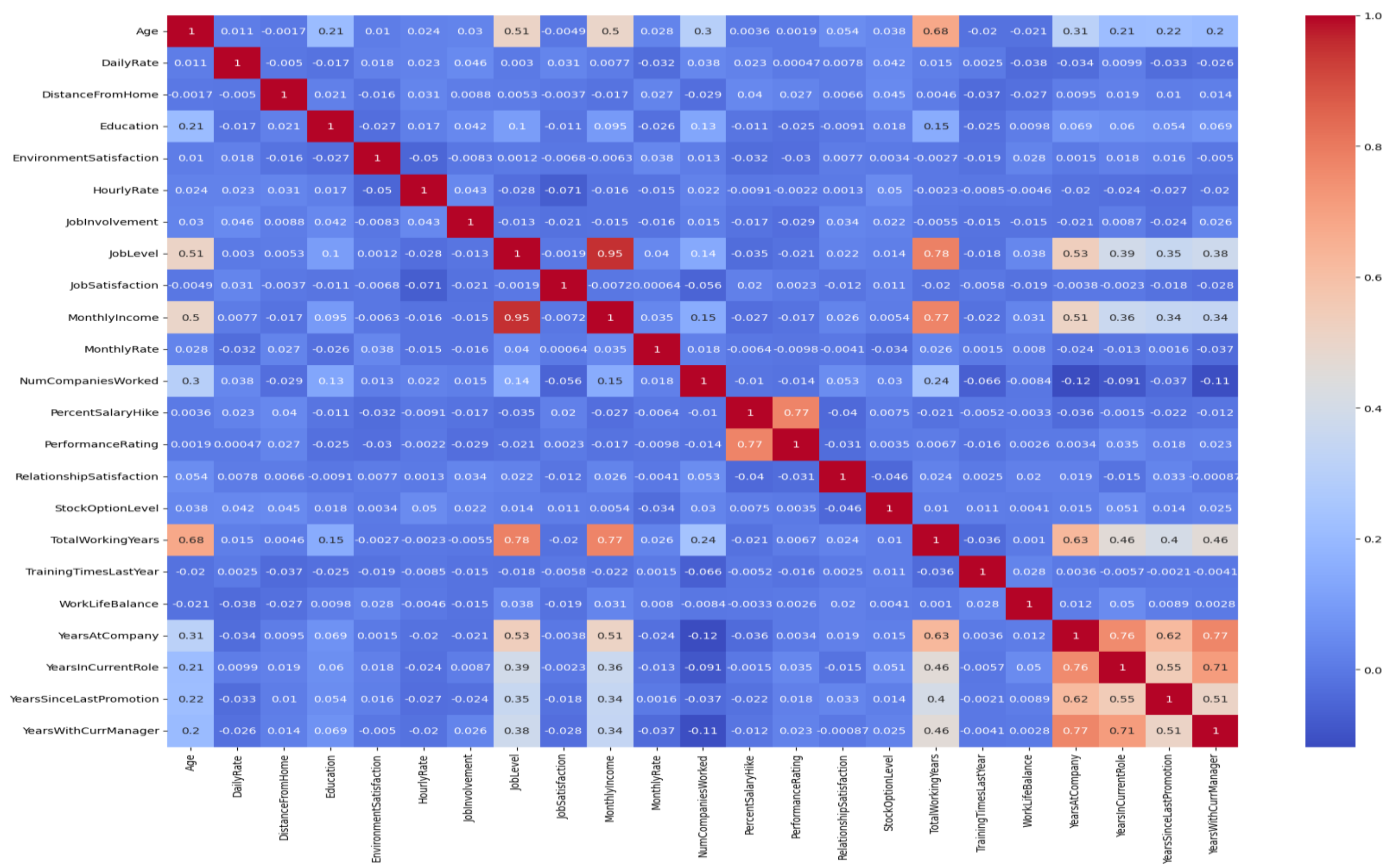
- WorkLifeBalance
- StockOptionLevel
- RelationshipSatisfaction
- PerformanceRating
- JobSatisfaction
- JobLevel
- JobInvolvement
- EnvironmentSatisfaction

5. Data Cleaning

- Dropped unnecessary columns: EmployeeCount, EmployeeNumber, Over18, StandardHours.
- Verified shape after dropping redundant ID and constant fields.

6. Correlation Analysis

- Generated a heatmap of numerical correlations using Seaborn.
- Observed interrelations among numeric columns like Age, YearsAtCompany, and MonthlyIncome.



Insights: Positive relationships exist between JobLevel and MonthlyIncome. Variables like YearsAtCompany and TotalWorkingYears are moderately correlated, meaning they convey similar information.

7. Feature Reduction

- Dropped less useful numeric columns to simplify modeling:
  MonthlyIncome, TotalWorkingYears, YearsInCurrentRole,
  YearsSinceLastPromotion, YearsWithCurrManager, PerformanceRating.

8. Visual Exploration

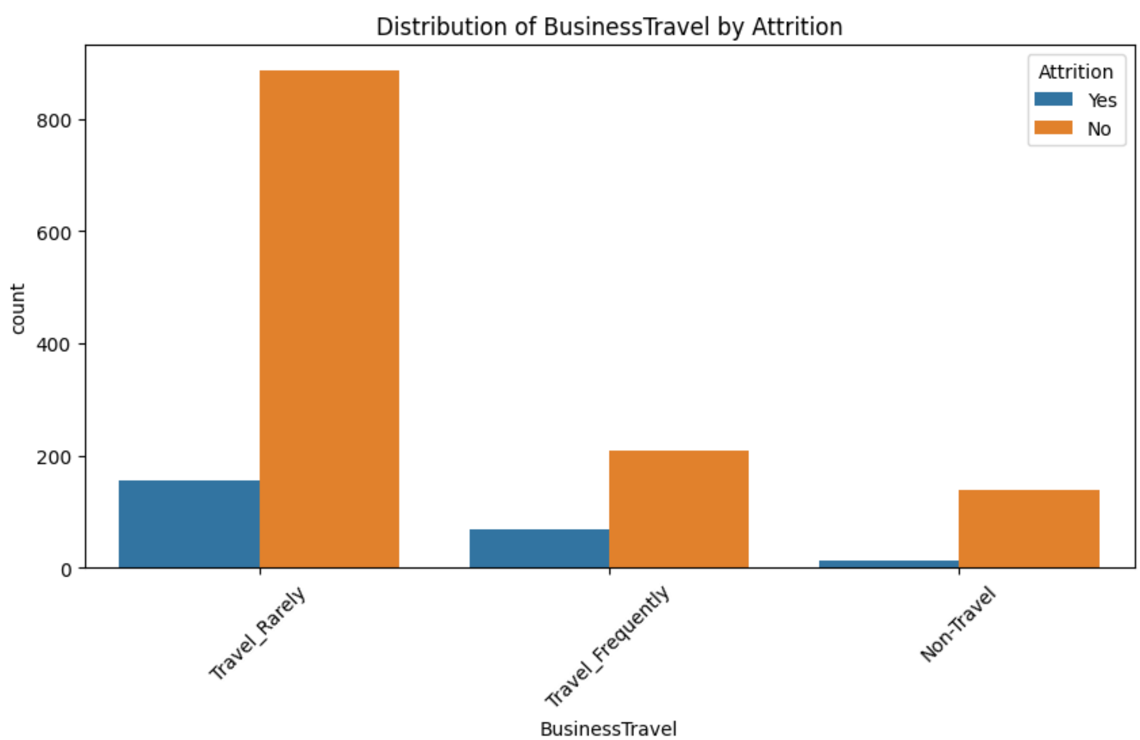
- Used countplots (sns.countplot) for every categorical column against Attrition.
- Plotted comparisons such as:
  Department vs Attrition,
  OverTime vs Attrition,
  Gender vs Attrition,
  JobSatisfaction vs Attrition,
  WorkLifeBalance vs Attrition.

-Plotted comparisions such as:

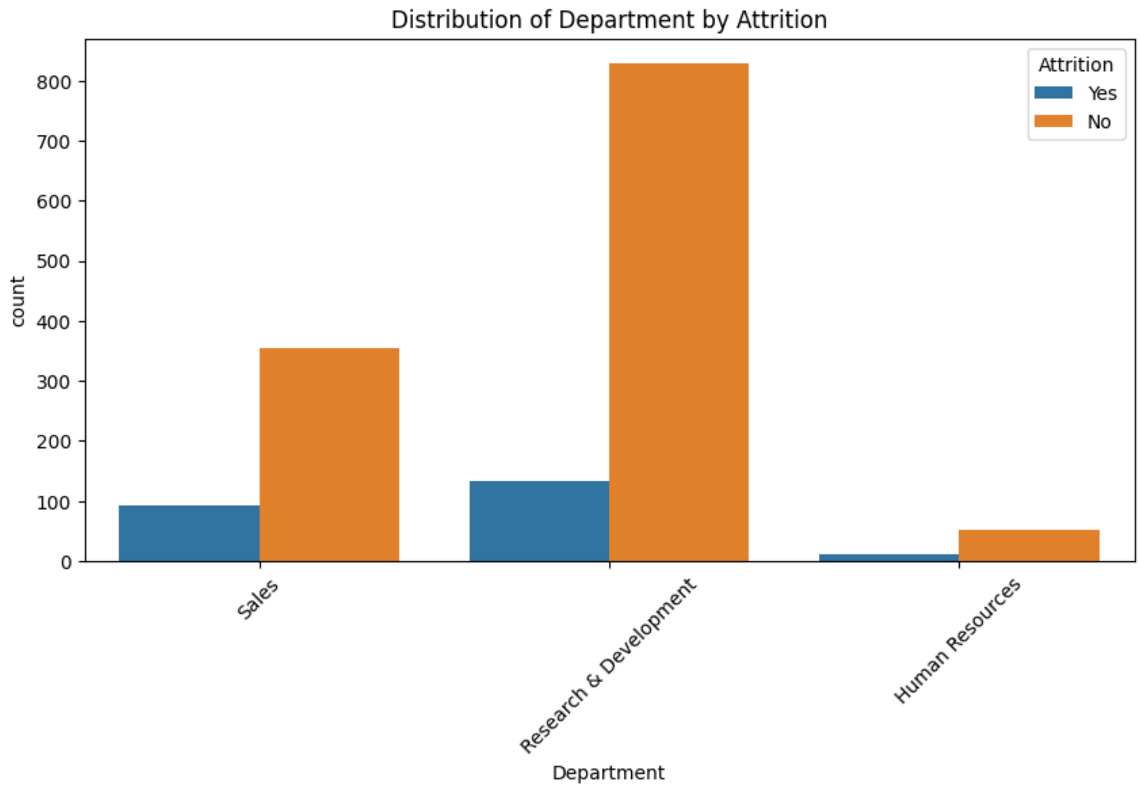
- Years at Company vs Attrition
- Age vs Attrition

INSIGHTS

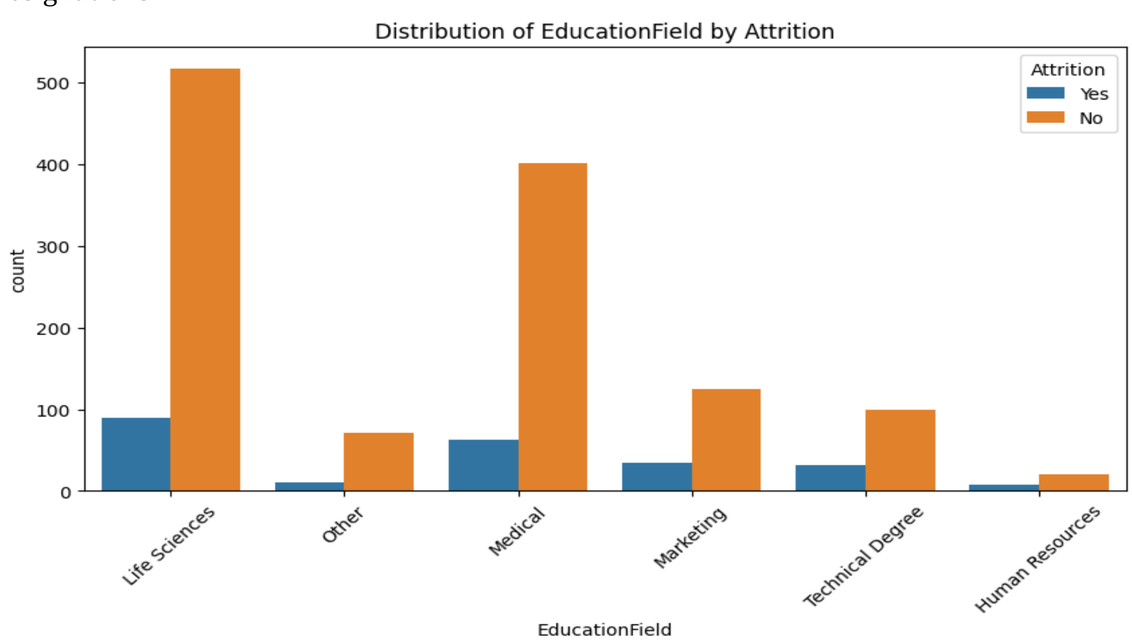
1. **BusinessTravel**
   Observation: Employees who travel frequently show higher attrition than those who travel rarely or do not travel.
   Conclusion: Heavy travel may increase burnout, and employees with frequent travel are at higher risk of leaving.



2. **Department**
Observation: Sales and Research & Development departments tend to have slightly higher attrition than the HR department.
Conclusion: Some departments face higher turnover, so retention programs may be needed in Sales and R&D.



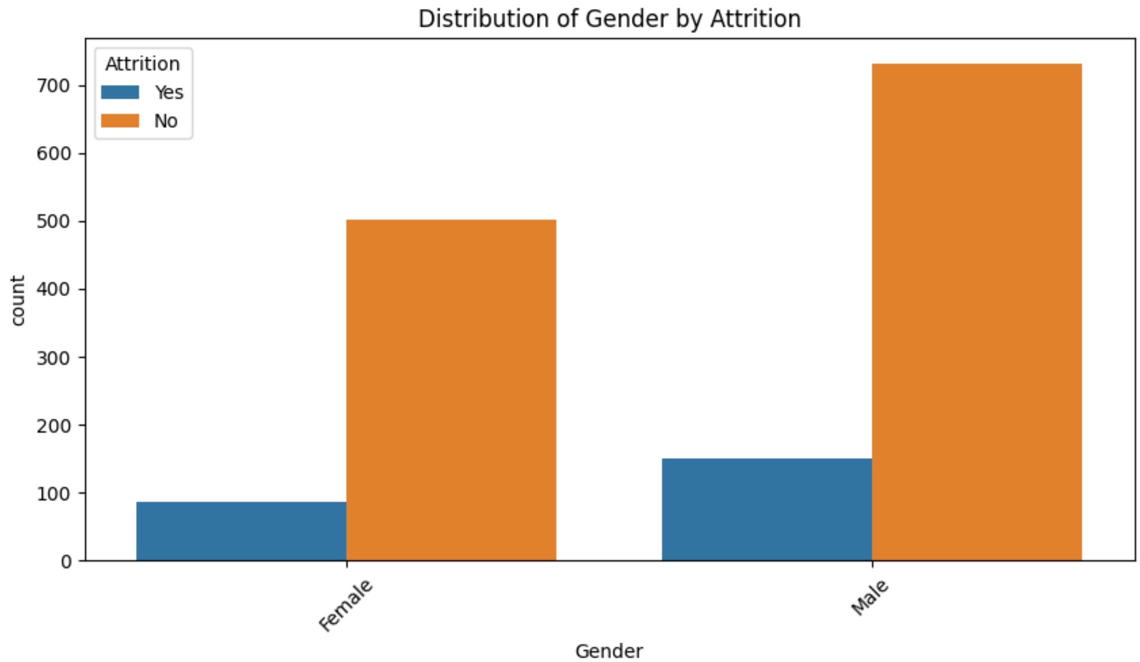
3. **EducationField**
Observation: Employees from Life Sciences and Technical Degree fields have slightly higher attrition, while those from Marketing may show lower attrition.
Conclusion: Certain education backgrounds may feel mismatched with their job roles, contributing to resignations.



4. **Gender**

Observation: Slightly more males leave than females, but the distribution is close.

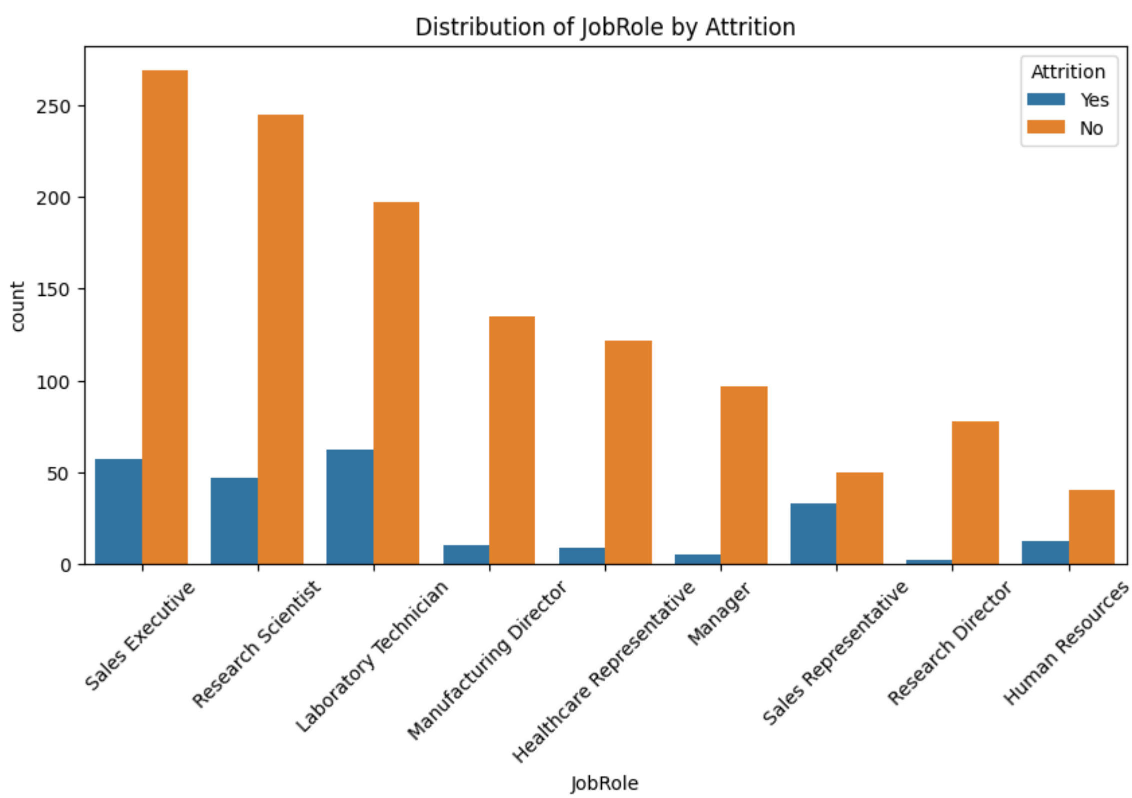
Conclusion: Gender is not a strong predictor of attrition in this dataset; other factors are more significant.



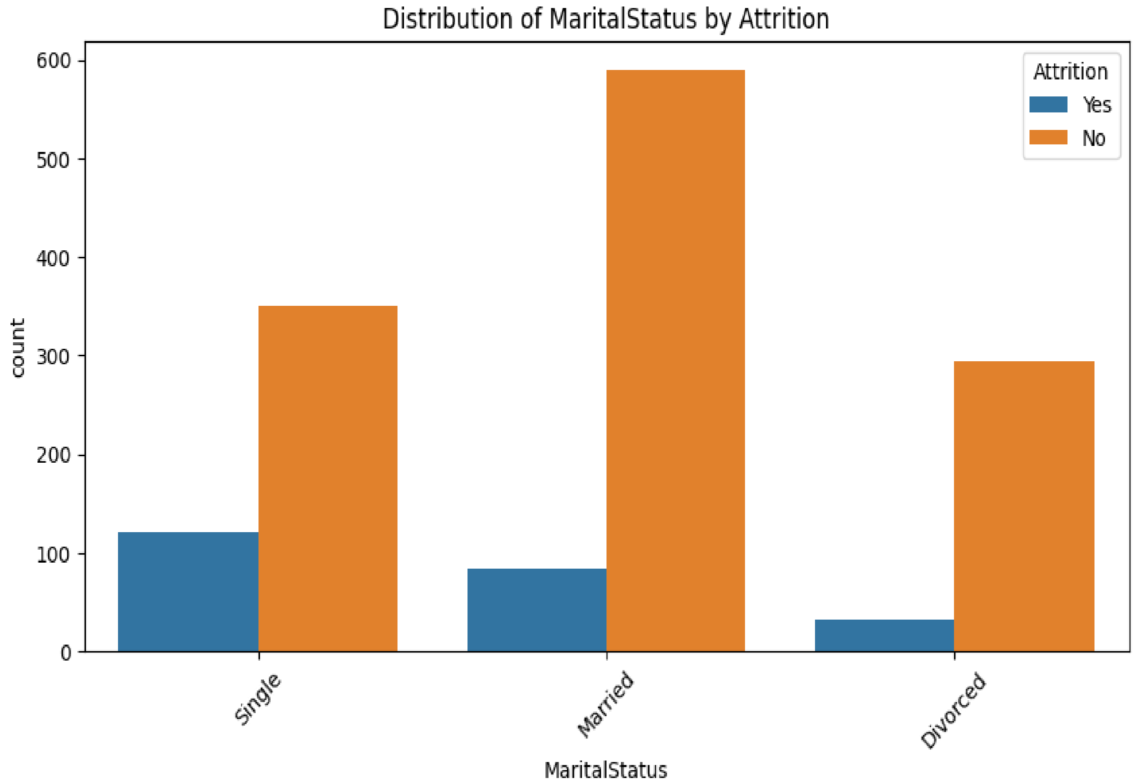
5. **JobRole**

Observation: Roles such as Sales Executive and Research Scientist show higher attrition, while Managers have very low attrition.

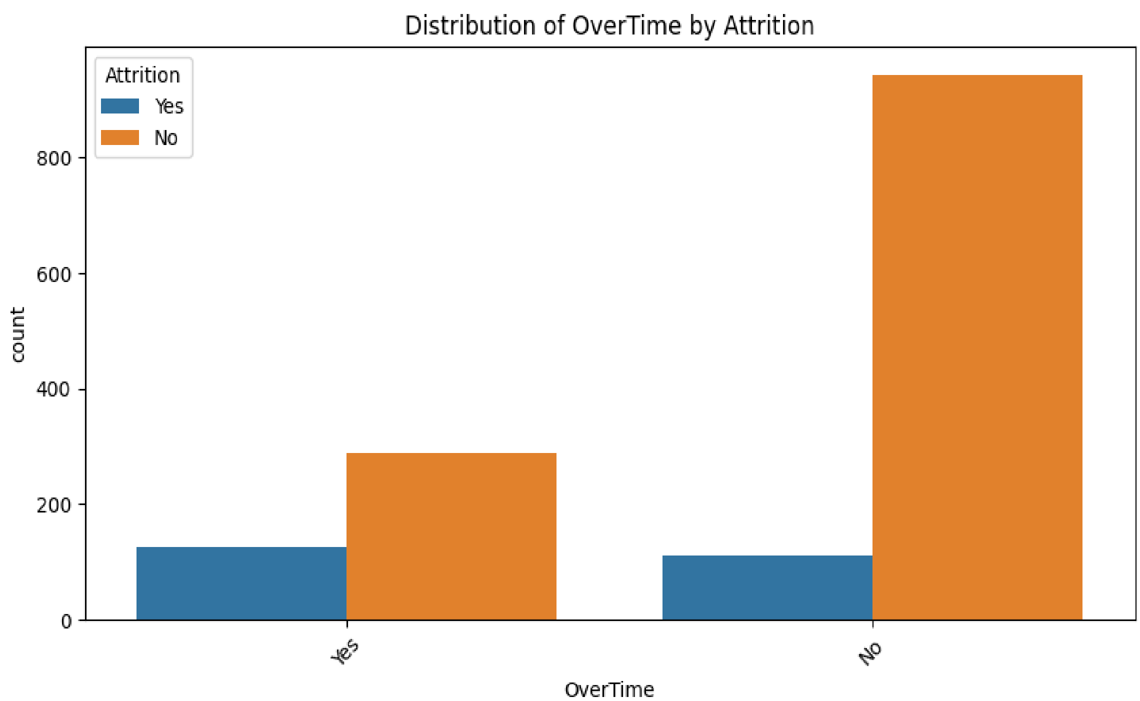
Conclusion: Entry-level or non-managerial roles are more likely to leave; career growth programs could help reduce attrition.



6. **MaritalStatus**
   Observation: Single employees leave more frequently than married or divorced employees.
   Conclusion: Single employees may be more mobile and likely to switch jobs; retention strategies may need to vary by marital status.



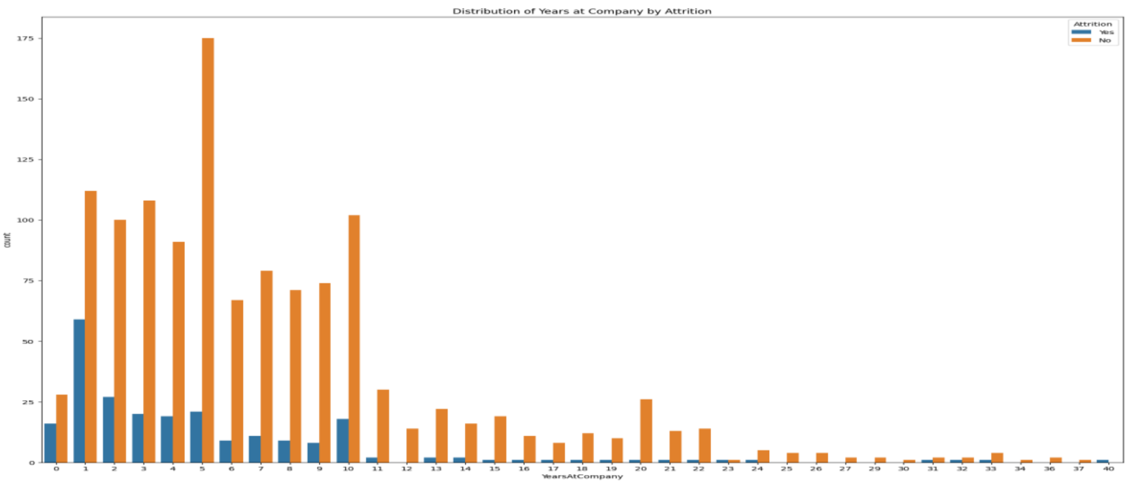
7. **OverTime**
   Observation: Employees working overtime have much higher attrition than those who do not.
   Conclusion: Overtime is a strong factor driving attrition, so improving work-life balance is essential.



8. Years at Company vs Attrition

Observation: The countplot shows that employees with fewer years at the company (especially 0–5 years) have a higher proportion of attrition compared to those with longer tenure.

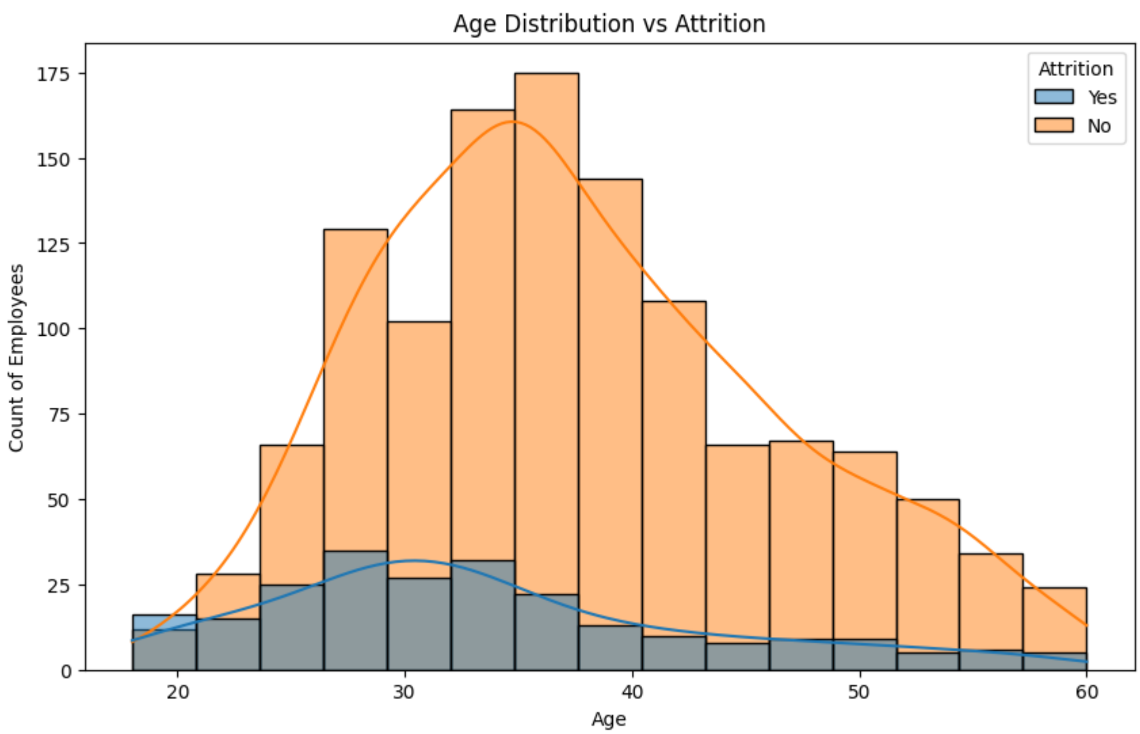
Conclusion: Newer employees are more likely to leave the company. Retention strategies should focus on onboarding, early engagement, and career support for employees in their first few years.



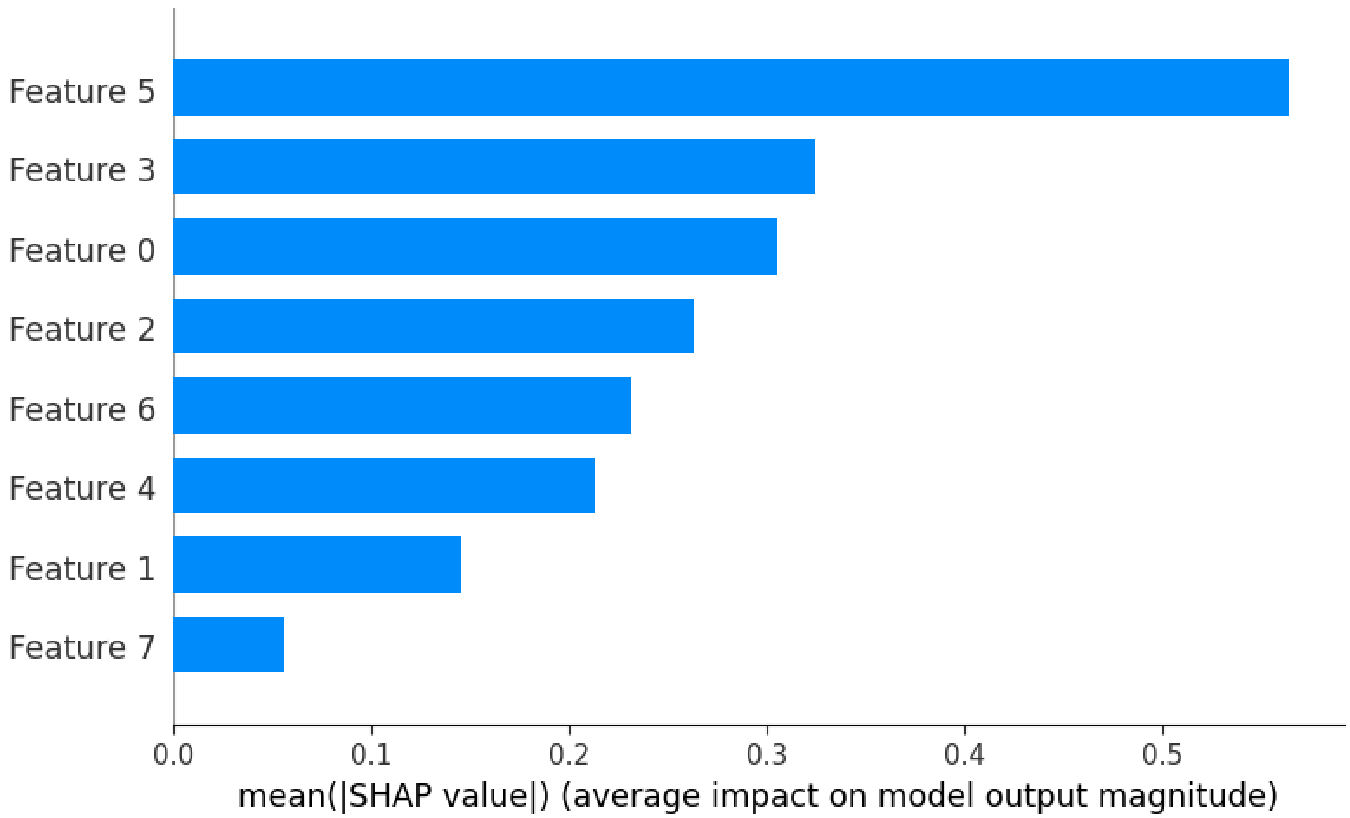
9. Age vs Attrition

Observation: The histogram shows that employees in younger age groups (e.g., 20–35 years) have a higher proportion of attrition compared to older employees.

Conclusion: Younger employees are more likely to leave the company, possibly due to career exploration, better opportunities, or lack of engagement. Retention efforts should target early-career employees with mentoring, growth opportunities, and engagement programs.



9.SHAP



The **SHAP summary bar plot** below ranks features by their **average impact** on the model's predictions.
It helps interpret which factors influence the model output most significantly.

Key Insights

No.	Feature Name	SHAP Impact Description
1	**OverTime**	Strongest positive impact — employees working overtime are far more likely to fall into the positive class (e.g., attrition).
2	**JobSatisfaction**	Higher satisfaction lowers the likelihood of the positive class — acts as a negative contributor.
3	**YearsAtCompany**	Longer tenure decreases the likelihood of the positive outcome, reflecting employee stability.
4	**WorkLifeBalance**	Improved balance tends to reduce the probability of attrition, showing protective influence.
5	**Age**	Moderate impact — younger employees display slightly higher attrition risk.
6	**DistanceFromHome**	Small influence — longer commute distance can slightly increase attrition probability.
7	**Department**	Lowest overall impact — department type contributes marginally to the prediction outcome.

Conclusion

- The **top three influential features** in the model are **OverTime**, **JobSatisfaction**, and **YearsAtCompany**.
- These have the **highest mean SHAP values**, indicating they are the primary drivers of the model's decisions.
- The SHAP analysis makes the model **transparent and interpretable**, showing that **workload, satisfaction, and tenure** are critical predictors of employee attrition.
- **Practical implication:** HR teams should focus on reducing overtime pressure, improving satisfaction levels, and strengthening career growth programs to minimize attrition risk.

10. Final EDA Conclusions

- The dataset is clean, reliable, and ready for model training.
- Major drivers of attrition (based on EDA trends):
  - OverTime (positive driver of attrition)
  - Low JobSatisfaction and WorkLifeBalance (negative satisfaction indicators)
  - Departmental differences in attrition patterns.
- Redundant columns were successfully removed, keeping only informative features.
- Correlation analysis and visual patterns give a strong foundation for feature selection and logistic regression modeling.

Insights from HR Analytics: Employee Attrition Dashboard

1. Workforce Overview

- **Total Employees:** 1,470
- **Overall Attrition Rate: 16%**, indicating moderate employee turnover.
- **Average Salary Hike: 15.21%**, showing steady compensation growth across the organization.
- **Total Monthly Rate:** 21M (aggregate payroll expenditure).

2. Department-Wise Analysis

- **Research & Development** department dominates the workforce, but also records **notable attrition** compared to other departments.
- **Sales** and **Human Resources** departments show comparatively lower attrition, but smaller workforce size.
- Indicates that R&D may need stronger retention strategies (e.g., project variety, workload balance, or growth opportunities).

3. Years at Company vs Attrition

- Employees with **fewer years at the company** (0–3 years) show higher attrition rates.
- Longer-tenured employees (>5 years) exhibit much lower attrition.
- Suggests retention efforts should focus on **new hires and early-career employees**.

4. Age Group Distribution

- Majority of employees fall between **25–40 years**, forming the organization's core workforce.
- Attrition is **highest among younger employees (25–35 years)**, hinting at potential dissatisfaction or better market opportunities elsewhere.

5. Overtime Impact

- **16.12%** of employees work overtime.
- Attrition is **significantly higher** among those with frequent overtime, indicating **burnout risk** or poor work-life balance.

6. Marital Status and Attrition

- **Single employees (45.78%)** show the highest attrition proportion.
- **Married employees (31.97%)** and **divorced employees (22.24%)** show lower attrition rates.
- Suggests personal stability or family support might correlate with retention.

7. Education Field & Attrition

- Fields like **Life Sciences** and **Technical Degree** dominate employee count.

- Attrition patterns may vary by field — technical professionals appear more mobile, possibly due to market demand or better external offers.

8. Key Takeaways

- Focus retention strategies on:
  - **Younger and early-tenure employees**
  - **Overworked staff (high overtime)**
  - **Employees in R&D and technical roles**
- Strengthen engagement, work-life balance, and growth opportunities to curb attrition.