Employee Attrition:
Data Exploration and Prediction

Ahmad Kurtubi

Outline

Executive Summary

Objectives and Business Questions

Dataset Overview

Methodology

Key Findings and Insights

Recommendations

Conclusion

Executive Summary

Project Objectives	• Identifying factors influencing employee attrition • Predicting which employees are likely to leave
Business Questions	• What factors are most strongly correlated with employee attrition? • How do work-life balance, job satisfaction, and environment satisfaction impact attrition? • Are demographic factors like age, marital status, or education field predictive of attrition? • Can we predict which employees are at risk of leaving the company?
Methodology	• Descriptive Statistics • EDA • Data Clustering using Elbow & K-Means • Data Imbalance Treatment using ADASYN • Prediction Model Random Forest • Model Evaluation using Confusion Matrix
Key Findings and Insights	• Demographic Factors: Age, marital status, and gender can influence attrition rates. Older employees and females tend to be more stable. • Job Role and Department: Certain roles, especially in sales and HR, have higher attrition rates. • Job Satisfaction and Work-Life Balance: Lower job satisfaction, environment satisfaction, and work-life balance are strongly correlated with higher attrition rates. • Income and Experience: Higher income and experience levels are associated with lower attrition rates, but this is not a universal trend. • Clustering Analysis: The data can be effectively clustered into two groups based on key features, revealing distinct patterns in employee behavior and attrition.
Recommendations	For Cluster 0: • Invest in Talent Development • Improve Work-Life Balance • Competitive Compensation • Mentorship and Coaching • Assign mentors to guide and support young employees. For Cluster 1: • Recognition and Rewards • Challenging Assignments • Succession Planning
Conclusion	By understanding the distinct characteristics of each cluster, the organization can implement targeted strategies to improve employee retention and satisfaction

Project Objectives

- Identifying factors influencing employee attrition.
- Predicting which employees are likely to leave.

The context of this project is Company XYZ (imaginary)

Business Questions

What factors are most strongly correlated with employee attrition?

- How do satisfaction scores relate to attrition?
- What roles or departments have the highest attrition rates?
- How does monthly income influence attrition?

How do work-life balance, job satisfaction, and environment satisfaction impact attrition?

- How does overtime work correlate with satisfaction and attrition?
- Are employees with lower work-life balance more likely to leave?
- How does job involvement relate to employee retention?

Are demographic factors like age, marital status, or education field predictive of attrition?

- Does marital status influence attrition rates?
- Are employees in specific education fields more likely to leave?
- How does age distribution differ for employees who left vs. stayed?
- Is there a difference in attrition rates between genders?

Can we predict which employees are at risk of leaving the company?

- What features provide the strongest predictive signal?
- How accurate is the predictive model for attrition?
- Can we identify high-risk employee groups for intervention?
- How does the model handle imbalanced classes in attrition prediction?

```
data.dtypes
```

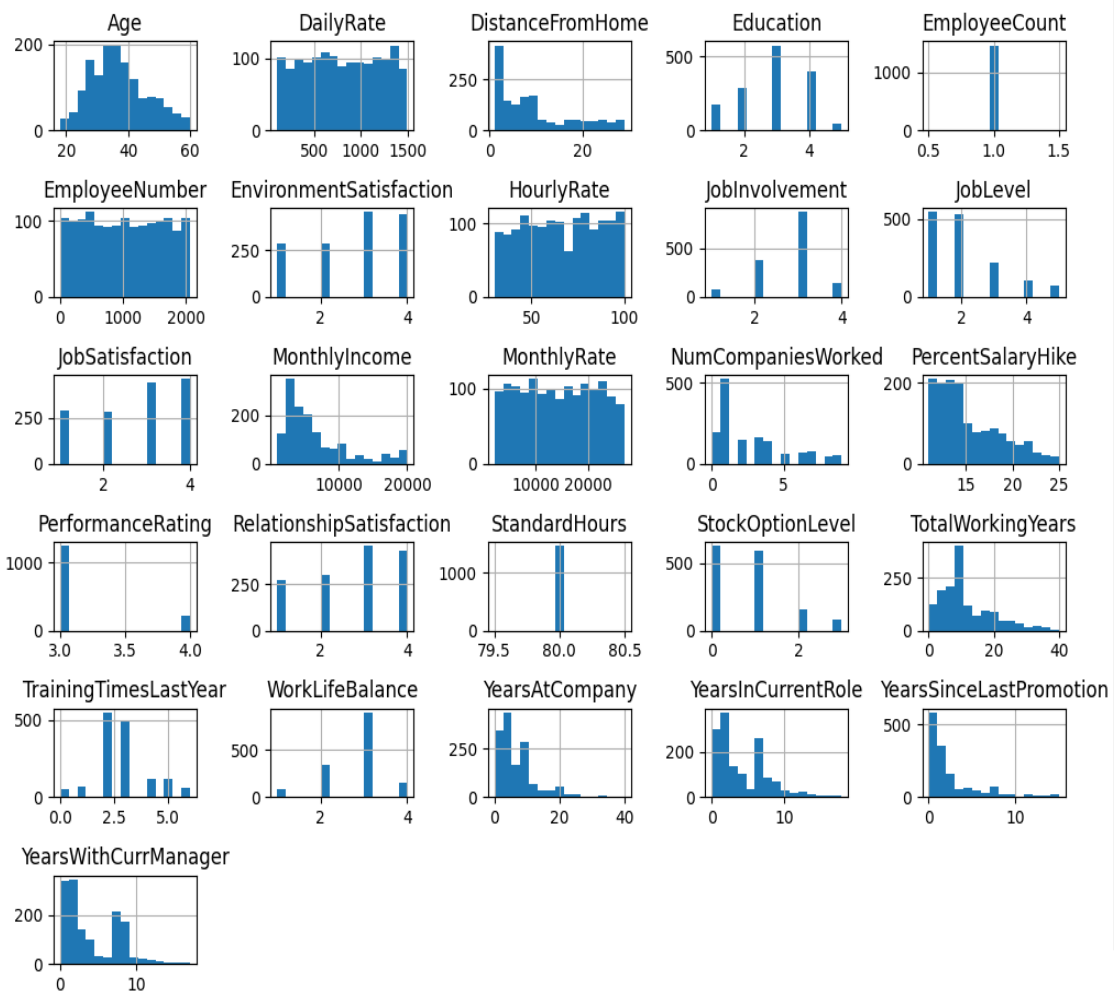
```
Age                         int
Attrition                   obj
BusinessTravel              obj
DailyRate                    in
Department                  obj
DistanceFromHome             in
Education                    in
EducationField              obj
EmployeeCount                in
EnvironmentSatisfaction      in
Gender                      obj
HourlyRate                   in
JobInvolvement               in
JobLevel                     in
JobRole                     obj
JobSatisfaction              in
MaritalStatus               obj
MonthlyIncome                in
MonthlyRate                  in
NumCompaniesWorked           in
OverTime                    obj
PercentSalaryHike            int
PerformanceRating            int
RelationshipSatisfaction     int
StandardHours                int
StockOptionLevel             int
TotalWorkingYears            int
TrainingTimesLastYear        int
WorkLifeBalance              int64
YearsAtCompany               int64
YearsInCurrentRole           int64
YearsSinceLastPromotion      int64
YearsWithCurrManager         int64
Attrition_Encoded            int64
OverTime_Encoded             int64
```

The Dataset Overview

- Dataset is synthetic
- Data license from IBM
- Downloadable in Kaggle

The Dataset Inspection

The Data Distribution: Initial Exploration and Insight



Age Distribution

- The majority of employees are between the ages of 30 and 40.
- There's a significant drop-off in employee numbers after the age of 40, suggesting potential age-related attrition factors.

Job Satisfaction & Environment Satisfaction

- The distribution of Job Satisfaction and Environment Satisfaction scores is relatively even across the range.
- However, a higher concentration of employees in the lower satisfaction levels might indicate potential areas for improvement to reduce attrition.

Work-Life Balance

- A significant portion of employees report a Work-Life Balance score of 3, suggesting a moderate level of balance.
- However, a non-negligible number of employees report lower scores, indicating potential stress and dissatisfaction that could contribute to attrition.

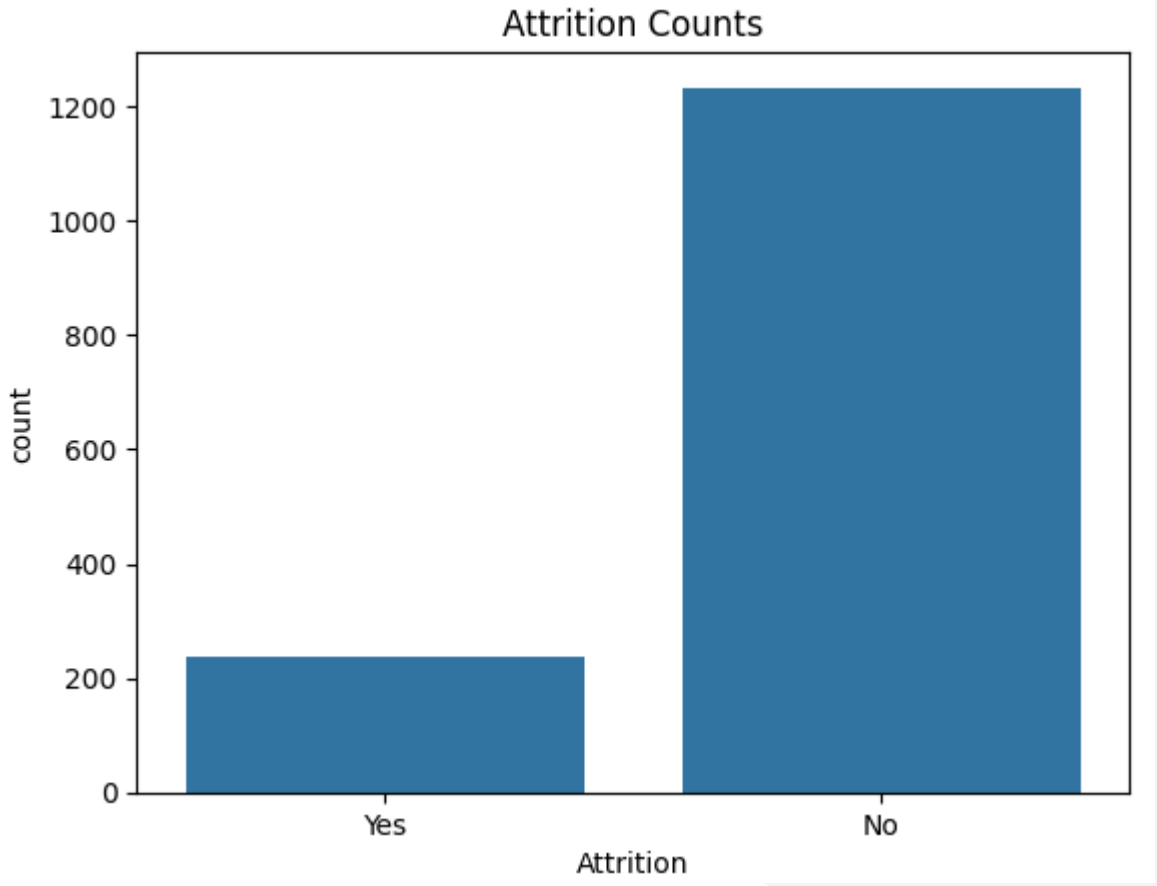
Job Involvement & Performance Rating

- The distribution of Job Involvement and Performance Rating scores is relatively even, suggesting a general level of engagement and performance among employees.

Tenure and Experience

- The distribution of years at the company, years in the current role, and years with the current manager shows a wide range of experience levels.
- A significant number of employees have been with the company for less than 5 years, suggesting potential turnover among newer employees.

The Data Distribution:
Initial Exploration and Insight



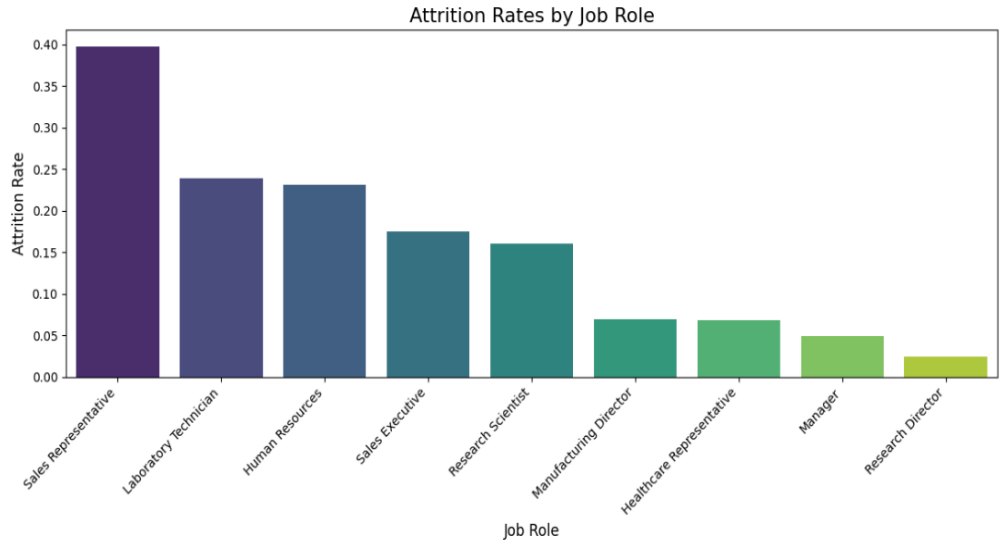
Distribution of Attrition: 16% Yes, 84% No.

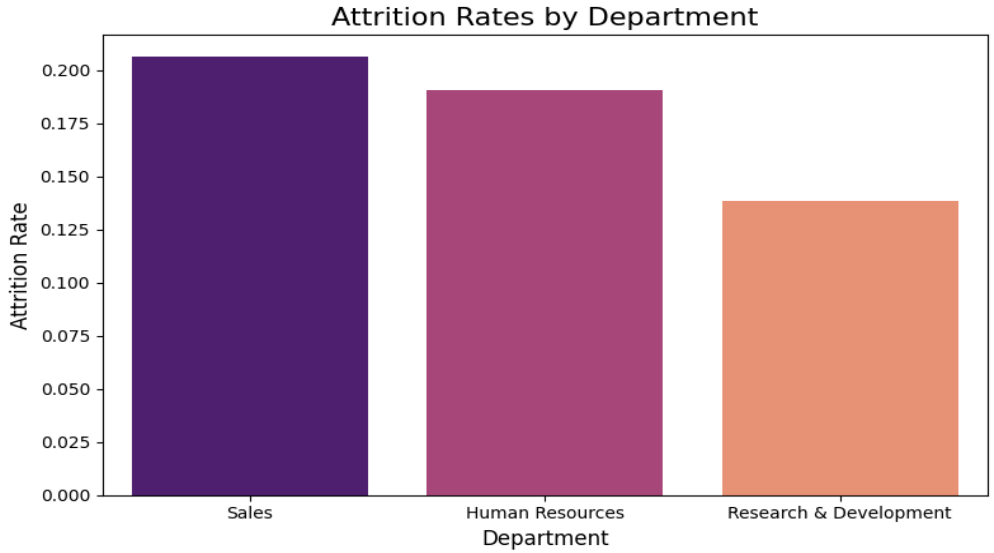
There is data imbalance in this data which potentially:

- **Brings bias toward the majority class** (“No Attrition / Staying in the Company”), leading to poor detection of the minority class (e.g., employees likely to leave).
- **Misleading Accuracy:** High accuracy can be misleading in imbalanced datasets. For example, if 90% of employees don’t leave, a model predicting "No" for every employee would achieve 90% accuracy but fail to identify employees at risk of leaving.

Exploratory Data Analysis (EDA)

• What Roles or Departments have the Highest Attrition Rates?





Department-Wise Attrition:

- **Sales Department:** This department has the highest attrition rate, indicating significant turnover.
- **Human Resources:** This department also has a relatively high attrition rate.
- **Research & Development:** This department has the lowest attrition rate, suggesting a more stable workforce.

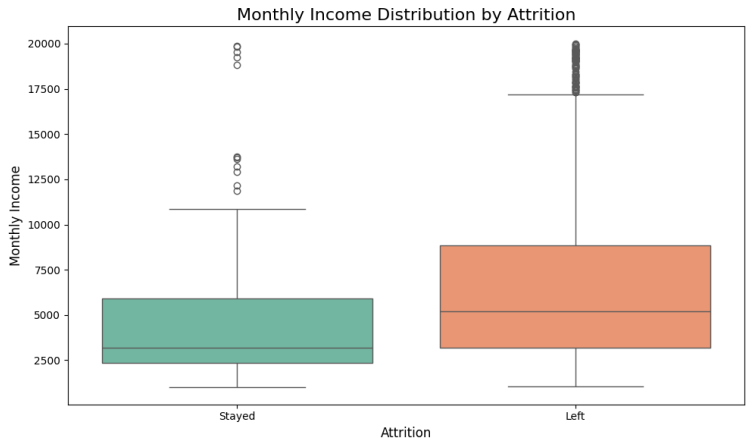
Role-Based Attrition:

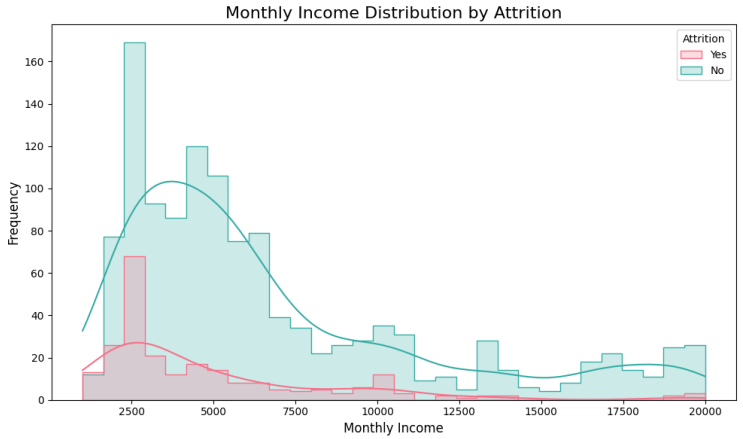
- **Sales Representative:** This specific role within the Sales department has the highest attrition rate.
- **Laboratory Technician and Human Resources:** These roles also have relatively high attrition rates.

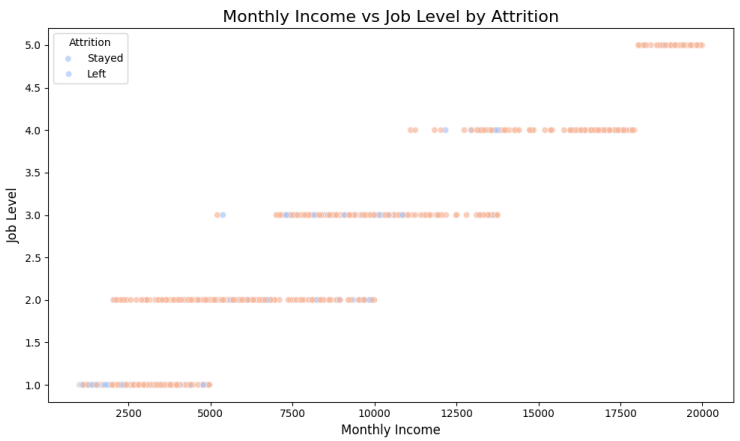
Potential Implications:

- **Targeted Retention Strategies:** Implement specific retention strategies for the Sales and Human Resources departments, especially for roles with high attrition rates.
- **Job Role Analysis:** Conduct a detailed analysis of job roles within these departments to identify factors contributing to high turnover.
- **Employee Satisfaction Surveys:** Conduct regular surveys to assess employee satisfaction and identify areas for improvement.
- **Career Development Opportunities:** Provide opportunities for career growth and advancement, especially for high-potential employees.
- **Work-Life Balance:** Implement policies and programs to improve work-life balance, such as flexible work arrangements and wellness programs.
- **Leadership Development:** Invest in leadership development programs to improve management practices and employee engagement.

• How does Monthly Income Influence Attrition?







Median Monthly Income:

- Employees who left the company tend to have a slightly higher median monthly income compared to those who stayed.

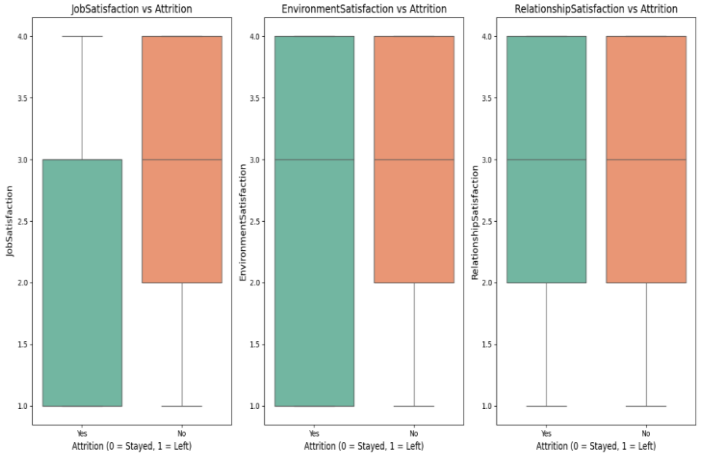
Income Distribution:

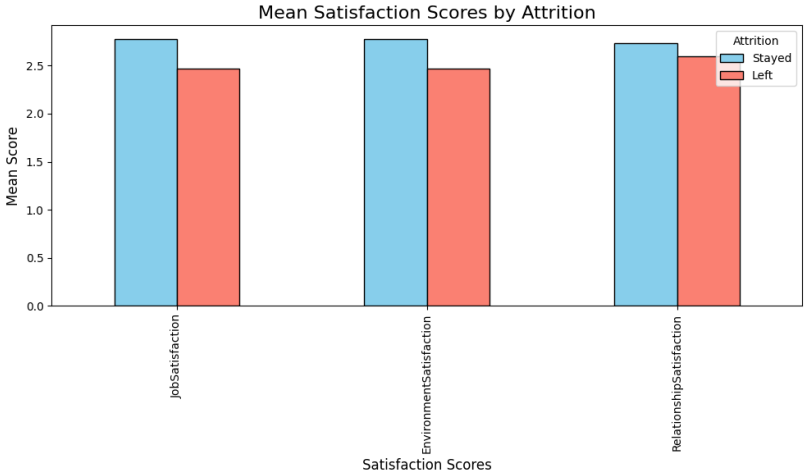
- Both groups exhibit a wide range of monthly incomes, indicating that income alone may not be a significant factor in attrition.

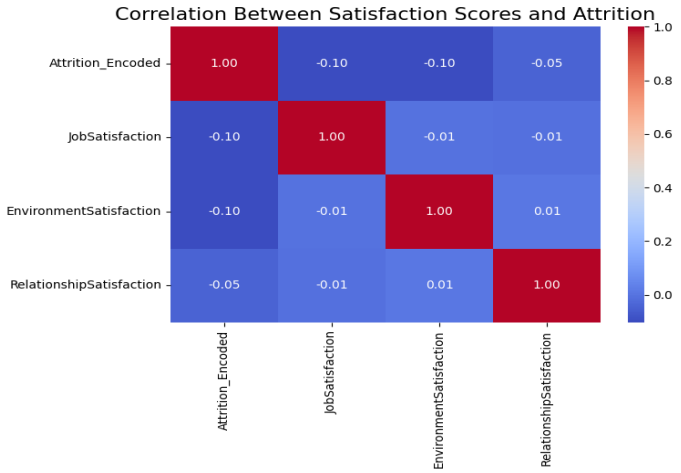
Outliers:

- There are a few outliers, especially among those who left, suggesting that some high-earning individuals may have left due to other factors, such as job dissatisfaction, lack of growth opportunities, or better offers.

• How do Satisfaction Scores Relate to Attrition?



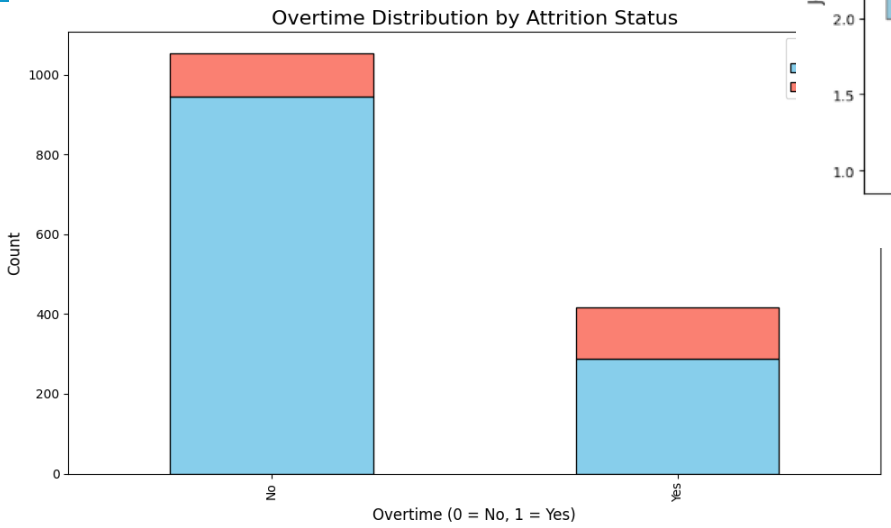


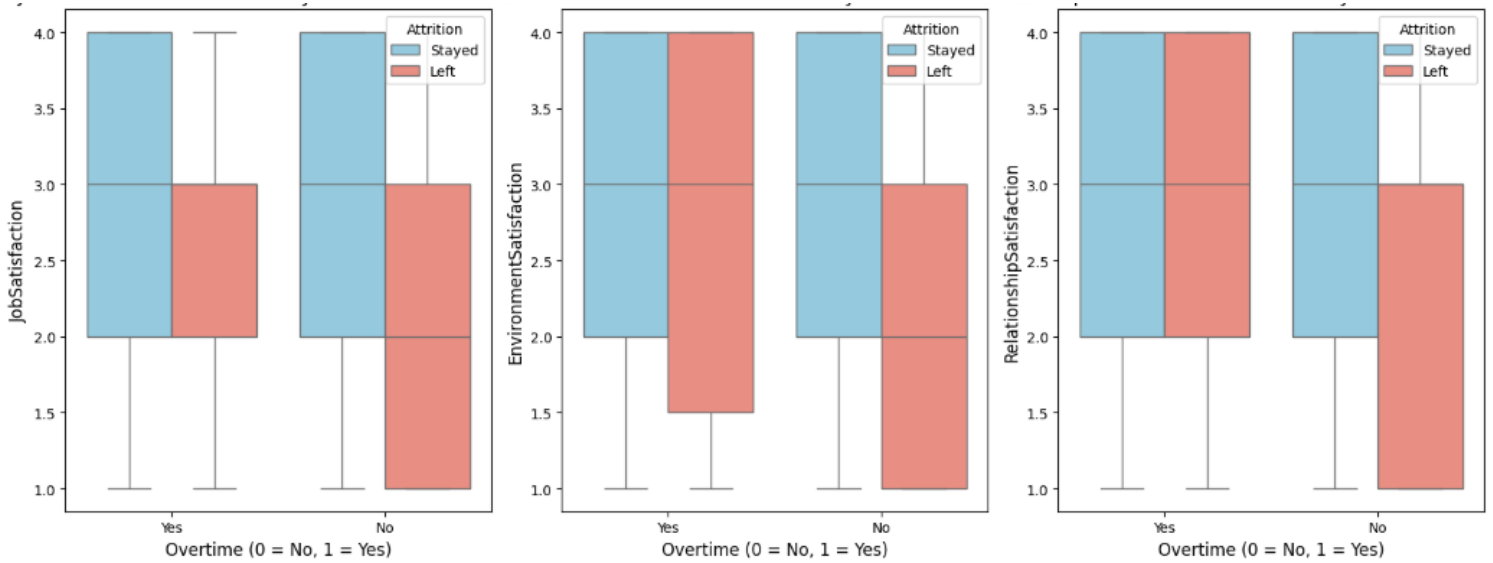


- **Lower Satisfaction, Higher Attrition**: Employees with lower job satisfaction, environment satisfaction, and relationship satisfaction are more likely to leave the company.

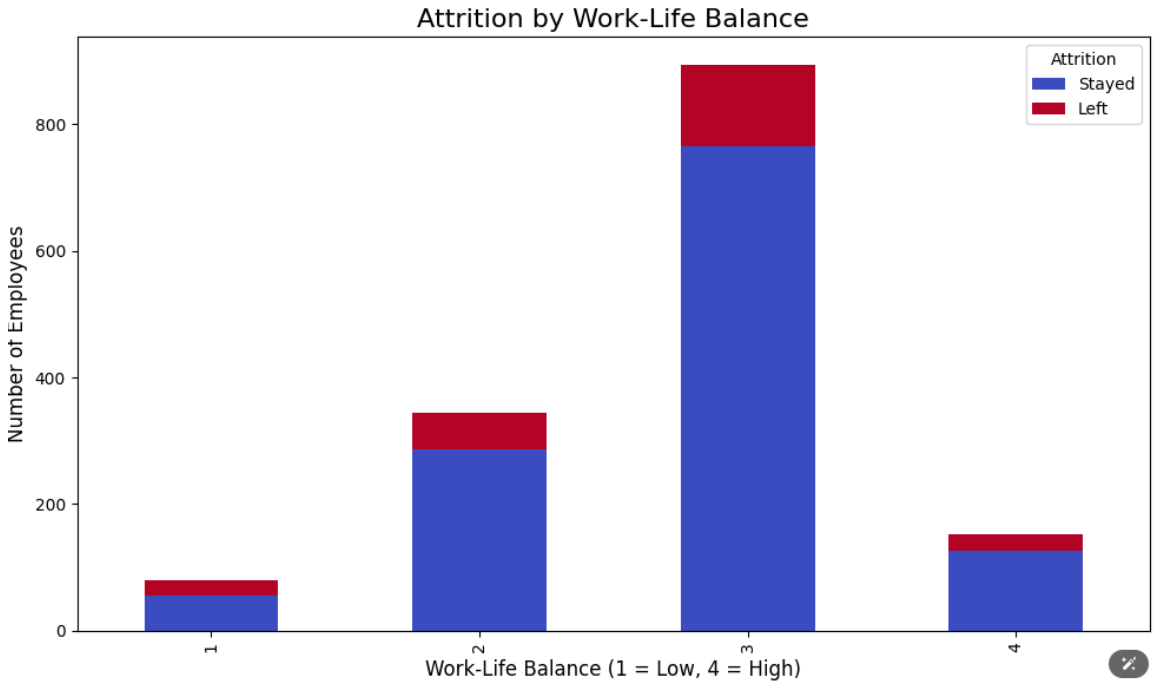
- **Variability in Satisfaction**: The wider range of satisfaction scores among those who left suggests that different factors might influence their decision to leave, beyond just dissatisfaction with their job or work environment.

- **How does Overtime Work Correlate with Attrition?**





- **Higher Attrition with Overtime:** Employees who work overtime are significantly more likely to leave the company compared to those who don’t.
- **Dominant Group:** Most employees who stayed did not work overtime.

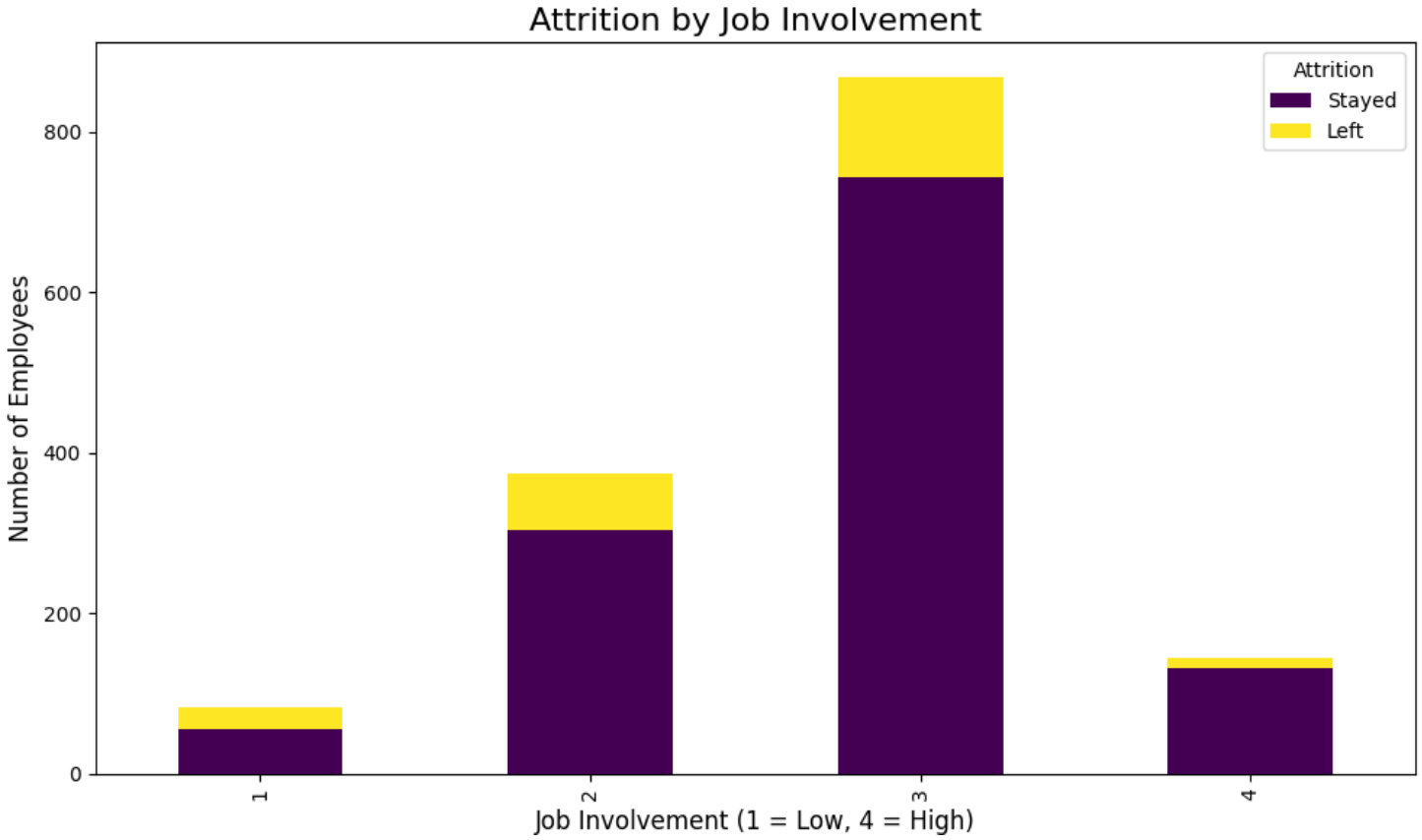


Are employees with lower work-life balance more likely to leave?

- Employees with a **higher work-life balance** (rating of 4) are more likely to **stay** with the company.
- Conversely, those with a **lower work-life balance** (ratings of 1, 2, and 3) are more likely to **leave** the company.

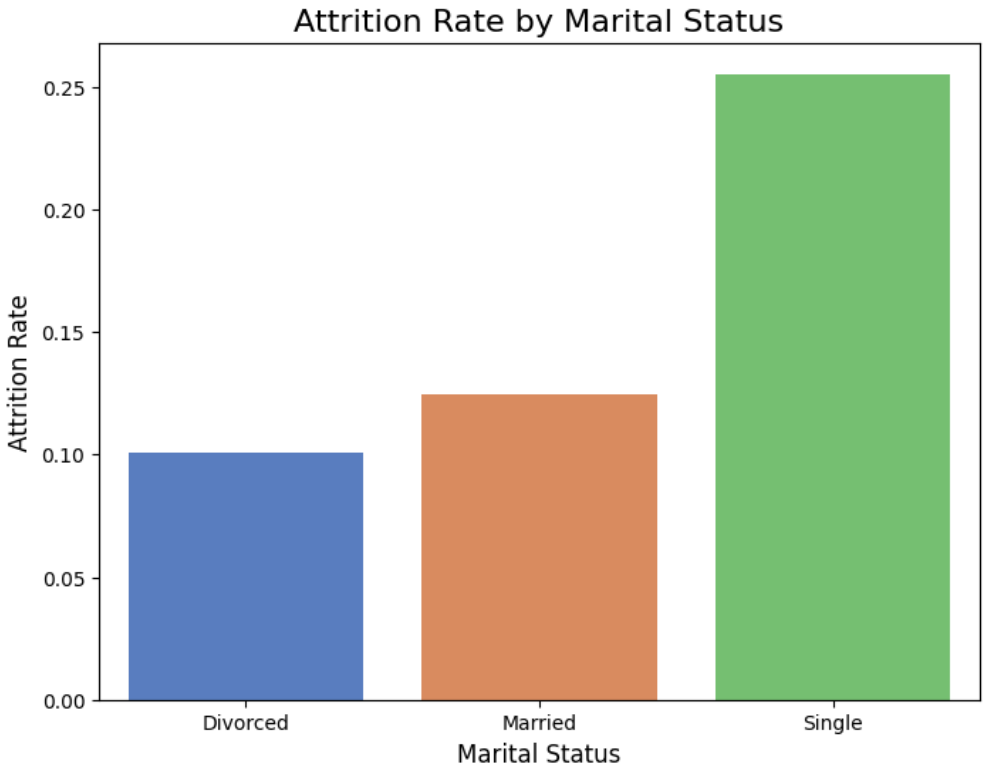
How does Job Involvement Relate to Employee Retention?

- Employees with **higher job involvement** (rating of 4) are more likely to **stay** with the company.
- Conversely, those with **lower job involvement** (ratings of 1, 2, and 3) are more likely to **leave** the company.

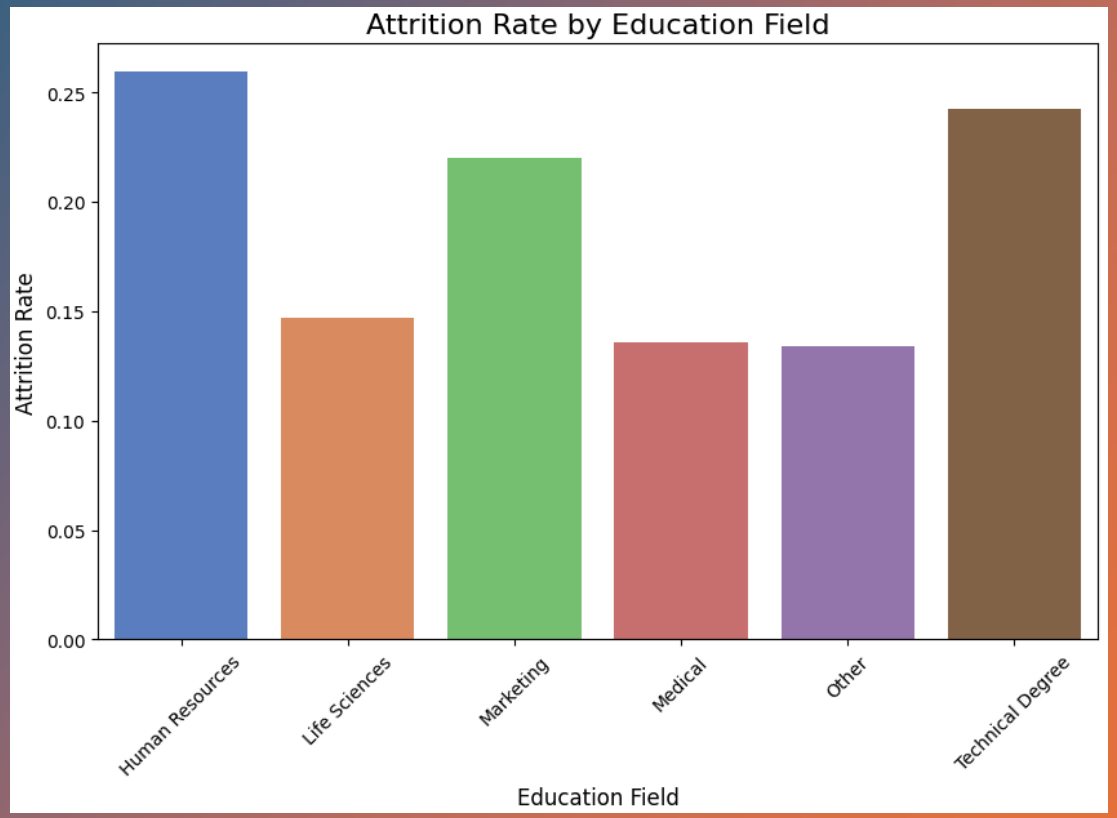


Does Marital Status Influence Attrition Rates?

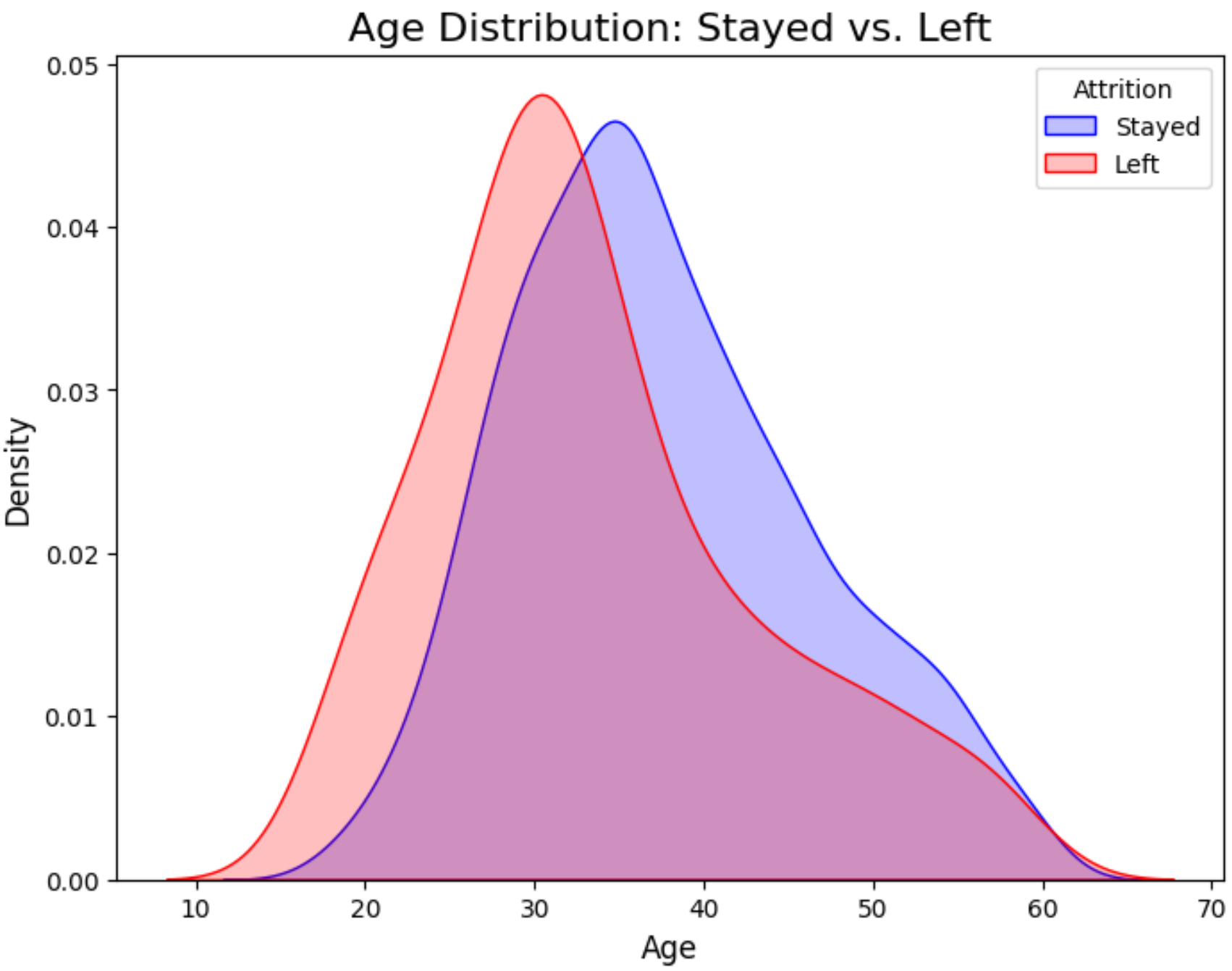
- **Single employees have the highest attrition rate**, followed by married employees, and then divorced employees.
- This suggests that marital status might be a factor influencing employee retention, with single employees being more likely to leave the company.



Are employees in specific education fields more likely to leave?

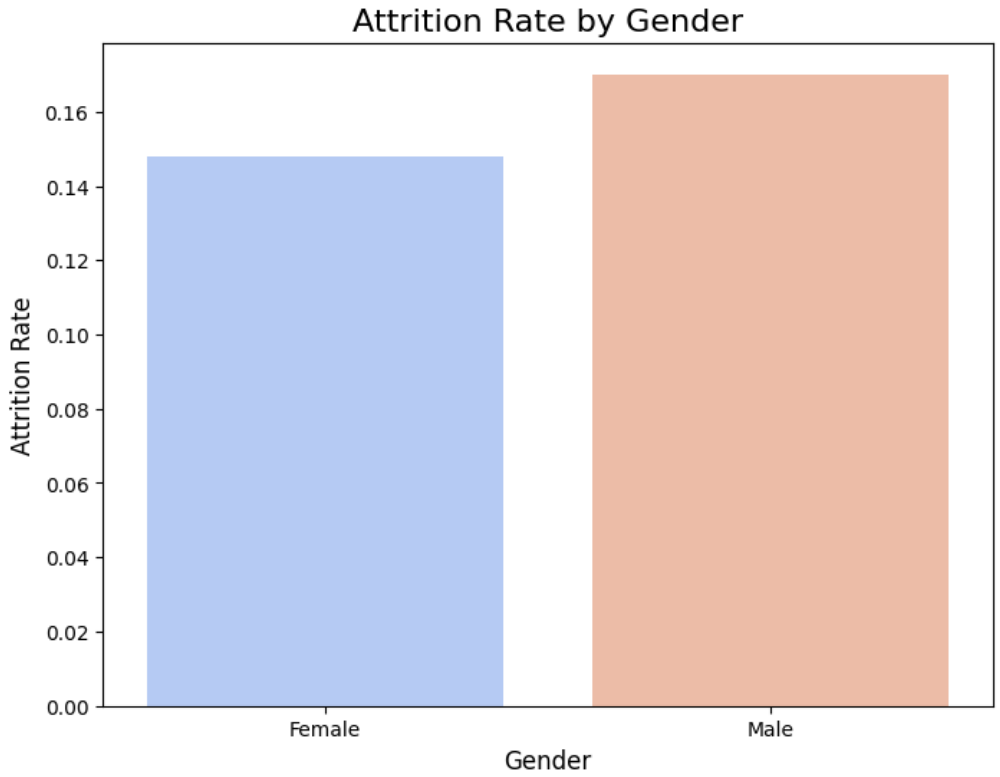


- Employees with a **Technical Degree** have the highest attrition rate, followed by those with a **Human Resources** background.
- This suggests that certain educational fields might be more prone to attrition than others.



How does Age Distribution Differ for Employees Who Left vs. Stayed?

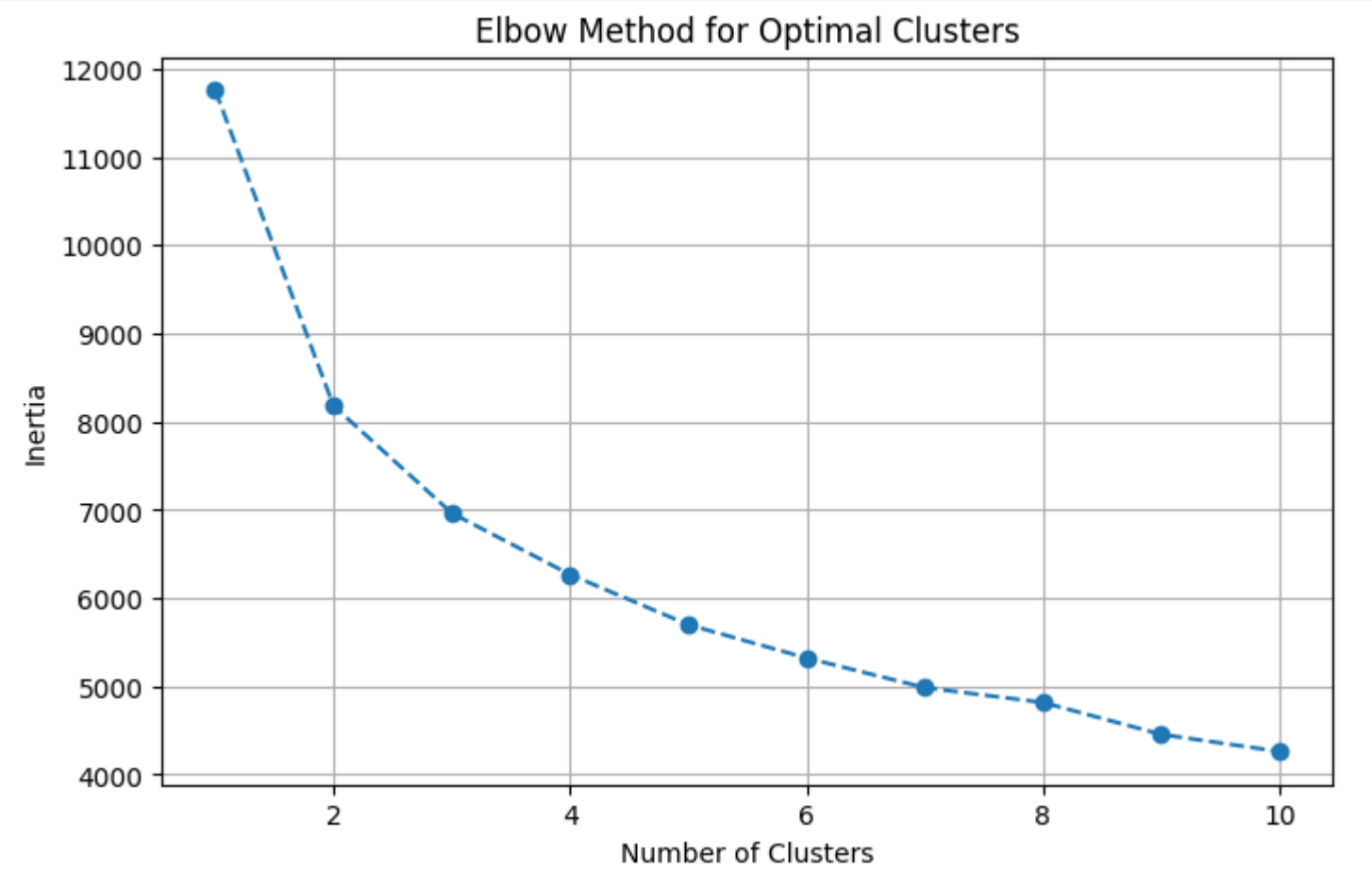
- Employees who **stayed** with the company tend to be **older** compared to those who left.
- The distribution of ages for employees who stayed is shifted slightly to the right, indicating a higher proportion of older employees in this group.



Is There a Difference in Attrition Rates between Genders?

- **Male employees have a higher attrition rate** compared to female employees.
- This suggests that gender might be a factor influencing employee retention, with male employees being more likely to leave the company.

Data Clustering



The Need for Data Clustering:

- Due to the variability of the pattern as depicted previously, there is a need to group the employees into clusters based on their respective similarities, and doing further analysis on each cluster afterwards

- Clustering will also help more targeted treatment based on the characteristics of the cluster

Chosen Number of Cluster:

- Based on **Elbow** method assessment, after 2 clusters, the significance of the differences is slowing down.

- Therefore, to see meaningful pattern, this project creates 2 clusters of employees for further analysis, using **K-Means**

Data Resampling to Handle Data Imbalance for Targeted Column “Attrition”

The resampling is conducted using ADASYN from imblearn.over_sampling library in Python

```
Processing Cluster 1...
Number of samples in original data: 286
Number of samples after resampling: 518
Number of positive (attrition) cases after resampling: 256
Number of negative (no attrition) cases after resampling: 262
```

```python
# Handle class imbalance using ADASYN
adasyn = ADASYN()
X_resampled, y_resampled = adasyn.fit_resample(X, y)

# Display sample sizes before and after resampling
print(f"Number of samples in original data: {len(cluster_data)}")
print(f"Number of samples after resampling: {len(X_resampled)}")
print(f"Number of positive (attrition) cases after resampling: {sum(y_resampled)}")
print(f"Number of negative (no attrition) cases after resampling: {len(y_resampled) - sum(y_resampled)}")
```

```
Processing Cluster 0...
Number of samples in original data: 1184
Number of samples after resampling: 1956
Number of positive (attrition) cases after resampling: 985
Number of negative (no attrition) cases after resampling: 971
```

Features (Columns) Used as Base for Clustering

Monthly Income :
- High importance and likely a strong differentiator for employees' financial satisfaction and stability.

Age:
- A demographic feature that influences career stage and priorities.

OverTime:
- A behavioral indicator that could reflect work habits or job demands.

Total Working Years:
- Captures overall experience, often linked to seniority and job stability.

Years At Company:
- Indicates tenure, which can highlight loyalty or the likelihood of attrition.

Distance From Home:
- A lifestyle factor that often correlates with satisfaction and likelihood of leaving.

Environment Satisfaction:
- Provides insight into how employees perceive their workplace.

Job Level:
- Reflects seniority, which can tie to compensation and responsibilities.

Rationale for Selection

These features are spread across key categories:

- **Demographic**:
  Age, Total Working Years
- **Compensation**:
  Monthly Income
- **Behavioral**:
  Over Time, Distance From Home
- **Satisfaction/Environment**: Environment Satisfaction
- **Job Role**:
  Job Level, Years At Company

The chosen features ensure diversity while avoiding redundancy (e.g., TotalWorkingYears vs. YearsInCurrentRole).

Resulted Clusters

Cluster 0:

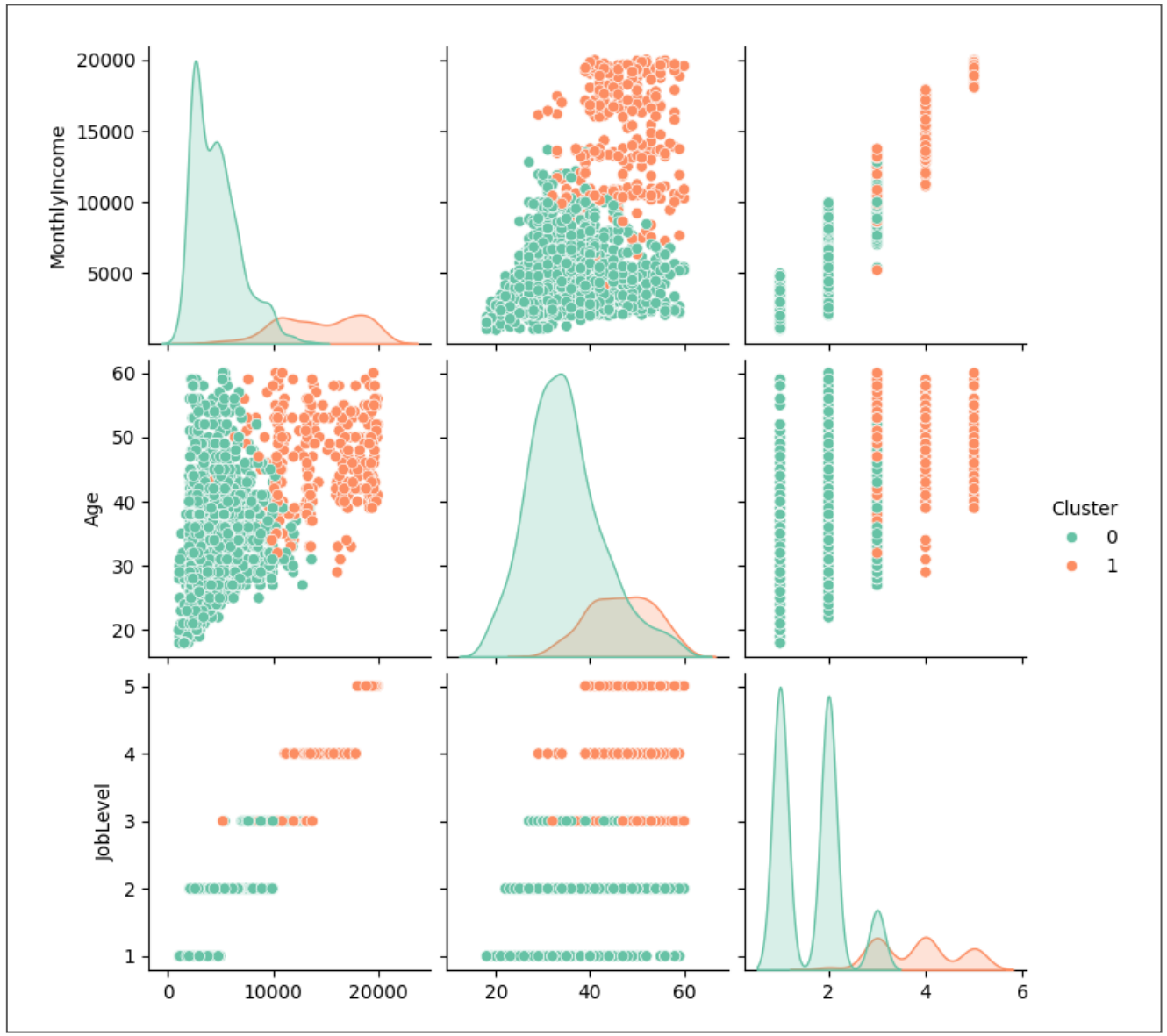
- **Younger Employees:** This cluster primarily consists of younger employees with lower monthly incomes and lower job levels.
- **Shorter Tenure:** They likely have fewer years of experience and a shorter tenure with the company.
- **Lower Job Involvement:** They might have lower levels of engagement and satisfaction with their work.

Cluster 1:

- **Older, More Experienced Employees:** This cluster primarily consists of older, more experienced employees with higher monthly incomes and higher job levels.
- **Longer Tenure:** They likely have a longer tenure with the company.
- **Higher Job Involvement:** They might have higher levels of engagement and satisfaction with their work.

Potential Implications for Attrition:

- **Cluster 0:** Employees in this cluster might be more prone to attrition due to lower job satisfaction, lack of career growth opportunities, or seeking higher-paying jobs.
- **Cluster 1:** Employees in this cluster might be more likely to stay with the company due to higher job satisfaction, strong relationships with colleagues, and career progression opportunities.



Prediction Model

Random Forest is chosen as prediction model algorithm for this project. Here's why:

- High Accuracy: It combines multiple decision trees, hence reducing overfitting and improves predictive accuracy.
- Handles Missing Values: It can handle datasets with missing values without requiring imputation.
- Feature Importance: It provides insights into the relative importance of different features in the prediction process. This feature importance is needed to answer the question related to influencing factors of attrition in this project.
- Robustness: It's robust to noisy data and outliers.
- Versatility: It can handle both numerical and categorical data.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.77      0.82       299
           1       0.79      0.88      0.83       288

    accuracy                           0.82       587
   macro avg       0.83      0.82      0.82       587
weighted avg       0.83      0.82      0.82       587


Confusion Matrix for Cluster 0:
              Predicted No Attrition  Predicted Attrition
No Attrition                     230                   69
Attrition                         35                  253
```

Performance of Prediction Model on Cluster 0

Overall Model Performance:

- The model achieves an **accuracy** of 82%, indicating that it correctly predicts whether an employee will stay or leave the company in 82% of cases.
- The **precision** and **recall** scores for both classes (0 and 1) are relatively balanced, suggesting that the model is reasonably good at both identifying true positives and true negatives.

Class-Specific Performance:

- **Class 0 (No Attrition):** The model has a higher precision (0.87) and recall (0.77), indicating that it's better at identifying employees who will stay with the company.
- **Class 1 (Attrition):** The model has a slightly lower precision (0.79) and a higher recall (0.88), suggesting that it's better at identifying employees who will leave the company, even if it might misclassify some as staying.

Confusion Matrix:

- **True Positives:** 230 employees who were predicted to stay and actually stayed.
- **True Negatives:** 253 employees who were predicted to leave and actually left.
- **False Positives:** 69 employees who were predicted to leave but actually stayed.
- **False Negatives:** 35 employees who were predicted to stay but actually left.

Overall, the model provides a reasonable level of accuracy in predicting employee attrition.

Performance of Prediction Model on Cluster 1

Overall Model Performance:

- The model achieves an accuracy of 86%, indicating that it correctly predicts whether an employee will stay or leave the company in 86% of cases.
- The precision and recall scores for both classes (0 and 1) are balanced, suggesting that the model is equally good at identifying both true positives and true negatives.

Class-Specific Performance:

- The model performs similarly well for both classes, with precision and recall scores of 0.86 for both.

Confusion Matrix:

- True Positives: 65 employees who were predicted to stay and actually stayed.
- True Negatives: 69 employees who were predicted to leave and actually left.
- False Positives: 11 employees who were predicted to leave but actually stayed.
- False Negatives: 11 employees who were predicted to stay but actually left.

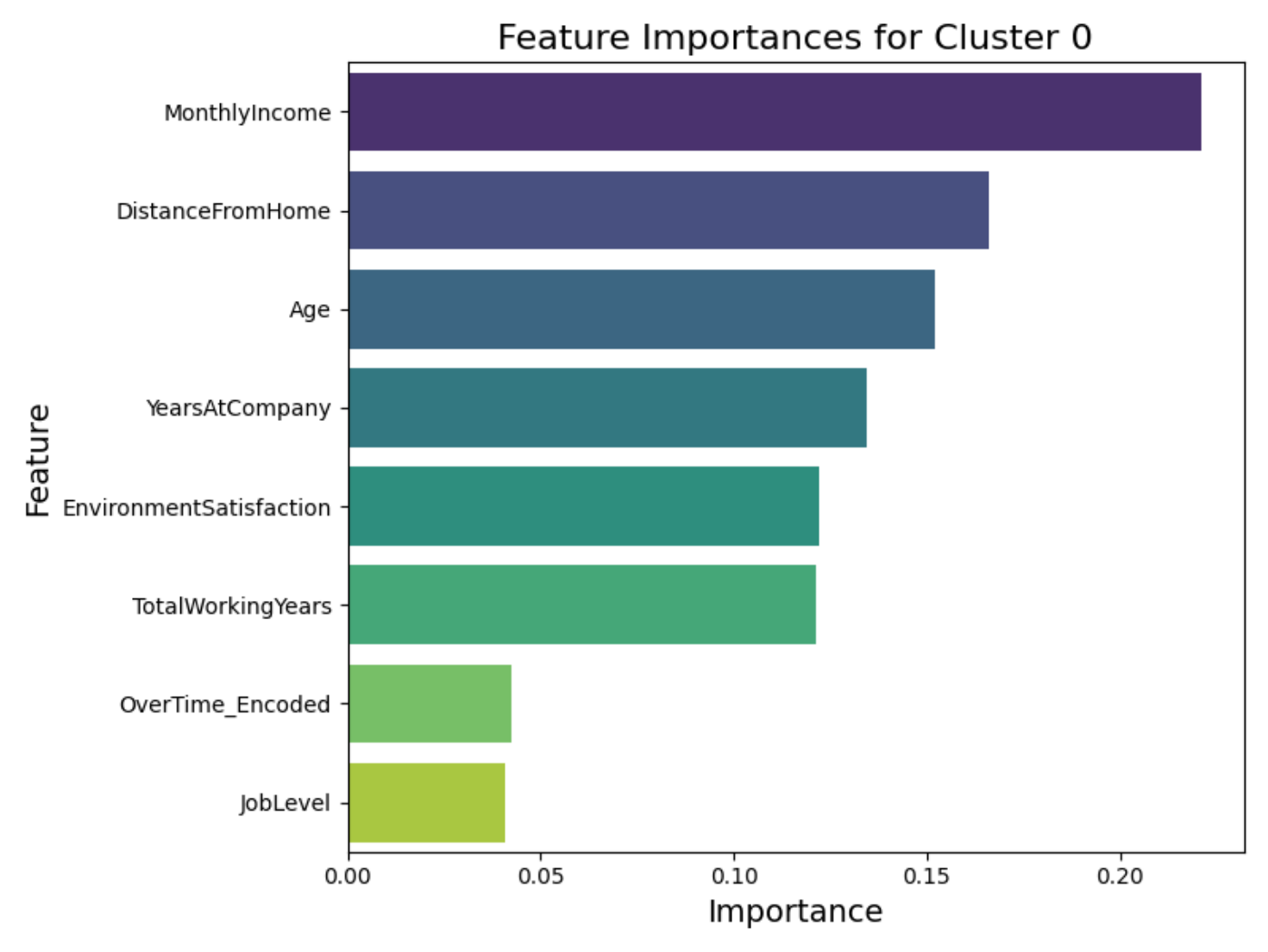
Overall, the model provides a strong performance in predicting employee attrition. It effectively balances precision and recall, making it reliable for identifying employees who are likely to leave the company.

Classification Report:

	precision	recall	f1-score	support
0	0.86	0.86	0.86	76
1	0.86	0.86	0.86	80
accuracy			0.86	156
macro avg	0.86	0.86	0.86	156
weighted avg	0.86	0.86	0.86	156

Confusion Matrix for Cluster 1:

	Predicted No Attrition	Predicted Attrition
No Attrition	65	11
Attrition	11	69



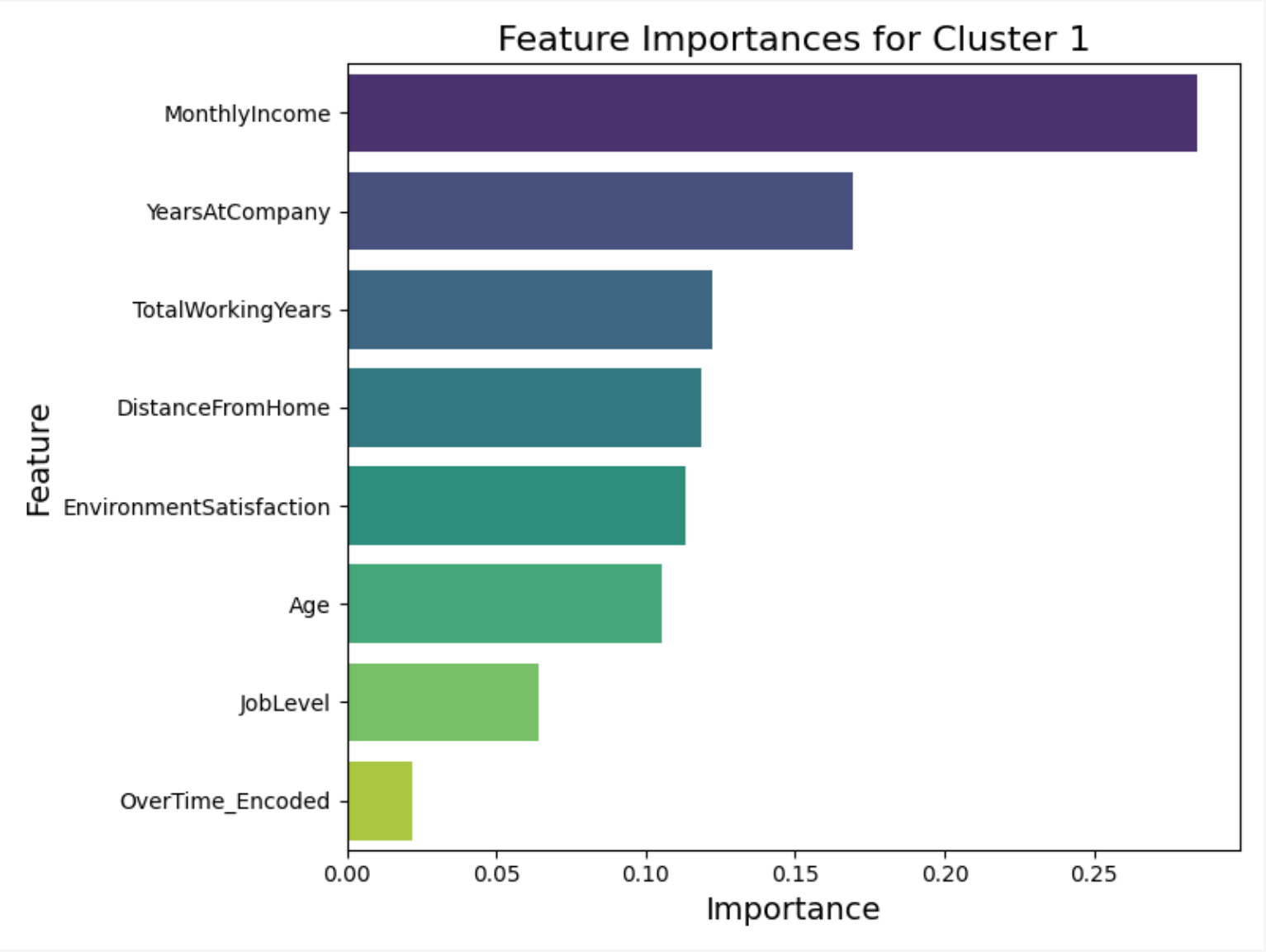
Most Influencing Factors of Attrition

For Cluster 0:

- **Monthly Income** is the most influential feature, suggesting that employees with similar income levels are grouped together.
- **Distance From Home** and **Age** also play significant roles in defining the cluster.
- Features like **Years At Company**, **Environment Satisfaction**, and **Total Working Years** have moderate importance.
- **Over Time** and **Job Level** have the least impact on cluster formation.

This suggests that the primary driver for clustering in this case is a combination of income, age, and proximity to work.

Most Influencing Factors of Attrition



For Cluster 1:

The chart shows that **Monthly Income** is the most important feature in determining cluster membership, followed by **Years At Company** and **Total Working Years**.

This suggests that employees in this cluster are likely to be older, more experienced, and have higher salaries. Factors like **Distance FromHome**, **Environment Satisfaction**, and **Age** also play a significant role, while **Over Time** and **Job Level** have a lesser impact.

Top 10 High–Attrition-Risk Employees

Cluster 0

Emp ID	Department	Job Level	Job Role	Age	Tenure (Y)
587	Research & Development	3	Healthcare Representative	31	1
1033	Research & Development	2	Healthcare Representative	37	1
811	Research & Development	1	Laboratory Technician	23	0
1052	Research & Development	2	Laboratory Technician	36	1
1107	Research & Development	1	Laboratory Technician	26	1
1156	Research & Development	1	Laboratory Technician	18	0
1604	Research & Development	1	Laboratory Technician	28	1
991	Research & Development	1	Research Scientist	31	3
952	Sales	1	Sales Representative	25	1
1273	Sales	1	Sales Representative	25	1

Top 10 High–Attrition-Risk Employees

Cluster 1

Employee ID	Department	Job Level	Job Role	Tenure (Y)
165	Research & Development	3	Healthcare Representative	40
582	Research & Development	3	Manager	7
825	Research & Development	5	Research Director	31
1572	Research & Development	3	Manufacturing Director	33
291	Sales	3	Sales Executive	14
307	Sales	3	Sales Executive	16
970	Sales	3	Sales Executive	14
1038	Sales	5	Manager	32
1639	Sales	3	Sales Executive	13
1968	Sales	3	Sales Executive	2

Key Findings and Implications

Cluster 0: Early Career, Lower Income

- **Younger Workforce:** This cluster primarily consists of younger employees.
- **Lower Job Satisfaction:** These employees tend to have lower job satisfaction and work-life balance.
- **Higher Attrition Risk:** This group is more likely to leave due to factors such as limited career growth opportunities and lower compensation.

Cluster 1: Experienced Professionals, Higher Income

- **Mature Workforce:** This cluster primarily consists of older, more experienced employees.
- **Higher Job Satisfaction:** These employees tend to have higher job satisfaction and work-life balance.
- **Lower Attrition Risk:** This group is less likely to leave due to factors like job security, higher compensation, and career stability.

Recommendations

For Cluster 0:

- **Invest in Talent Development:** Provide training and development opportunities to enhance their skills and prepare them for career advancement.
- **Improve Work-Life Balance:** Implement flexible work arrangements and wellness programs to reduce stress and burnout.
- **Competitive Compensation:** Review compensation packages to ensure they are competitive and aligned with industry standards.
- **Mentorship and Coaching:** Assign mentors to guide and support young employees.

For Cluster 1:

- **Recognition and Rewards:** Implement recognition programs to appreciate their contributions.
- **Challenging Assignments:** Provide opportunities for career advancement and leadership roles.
- **Succession Planning:** Identify and develop high-potential employees to ensure a smooth transition and continuity.

Conclusion

- With efforts in understanding the distinct characteristics of each cluster, the organization can implement targeted strategies to improve employee retention and satisfaction.
- By addressing the specific needs of each group, the organization can create a more engaged and productive workforce.

Appendices

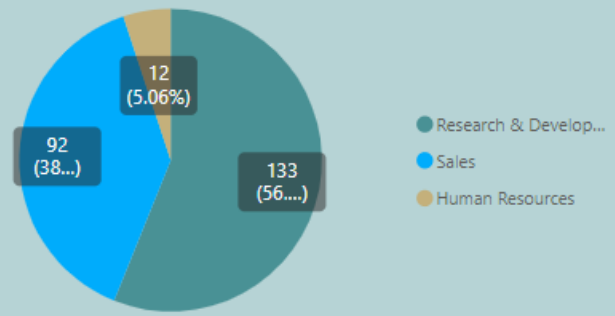
- Project assets (dataset, python files, etc) are accessible through the following github link
- The screenshots of interactive dashboard can be found at the end part of this presentation file

EMPLOYEE ATTRITION DASHBOARD

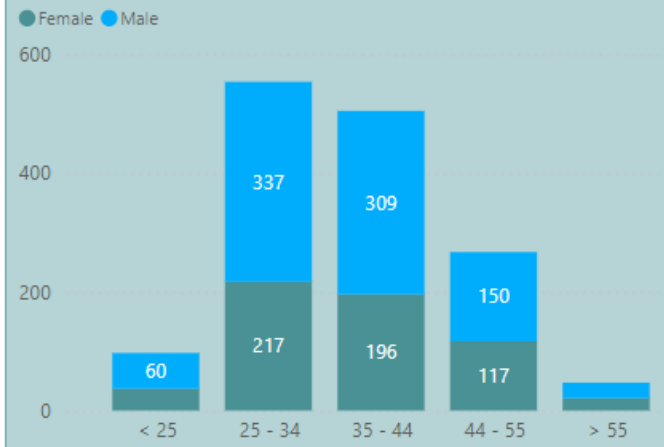
Basic Facts | Data Exploration | Prediction

Number of Employees	Number of Employees	Attrition Rate	Active Employees	Average Age
1470	237	16.1%	1233	37

Attrition by Department



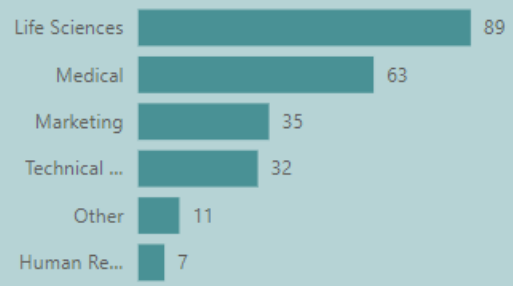
Attrition by Age Range



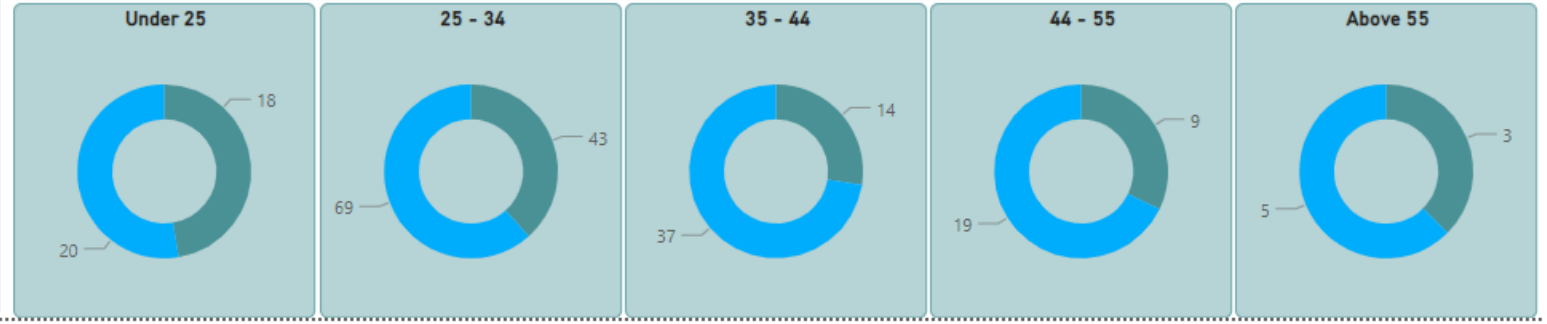
Job Satisfaction Score

JobRole	1	2	3	4	Total
Sales Representative	12	21	27	23	**83**
Sales Executive	69	54	91	112	**326**
Research Scientist	54	53	90	95	**292**
Research Director	15	16	27	22	**80**
Manufacturing Director	26	32	49	38	**145**
Manager	21	21	27	33	**102**
Laboratory Technician	56	48	75	80	**259**
Human Resources	10	16	13	13	**52**
Healthcare Representative	26	19	43	43	**131**
Total	**289**	**280**	**442**	**459**	**1470**

Attrition by Education Field

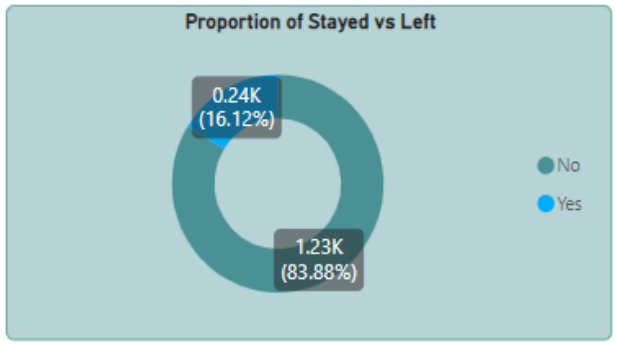


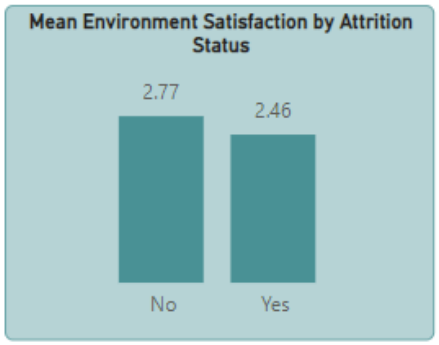
Attrition Rate by Age Range

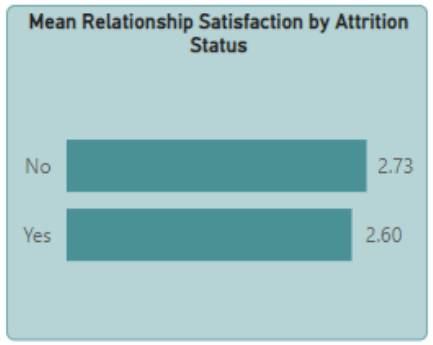


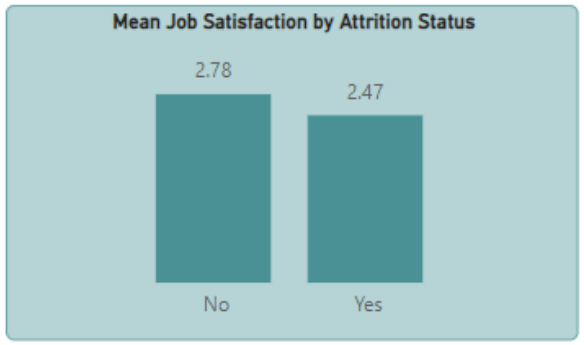
EMPLOYEE ATTRITION DASHBOARD

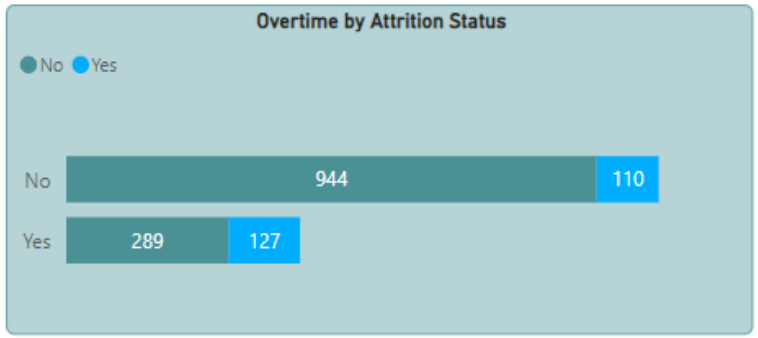
Basic Facts | Data Exploration | Prediction

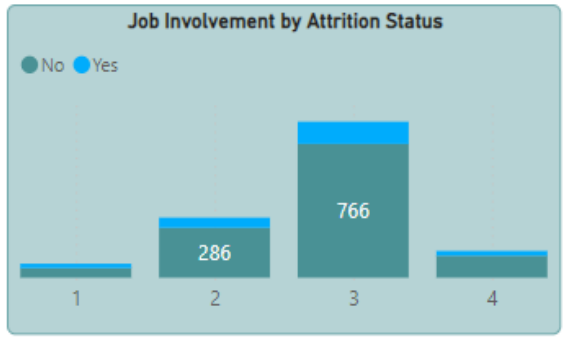


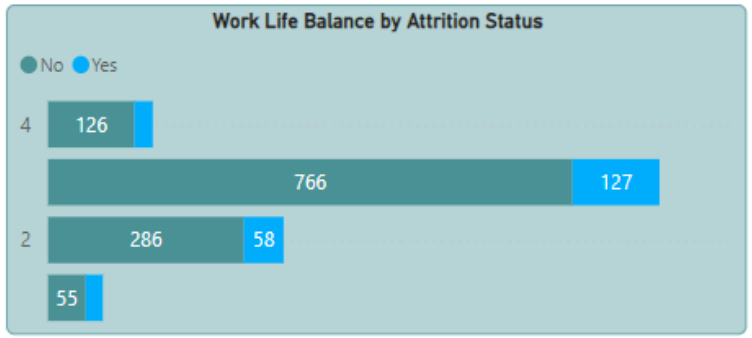


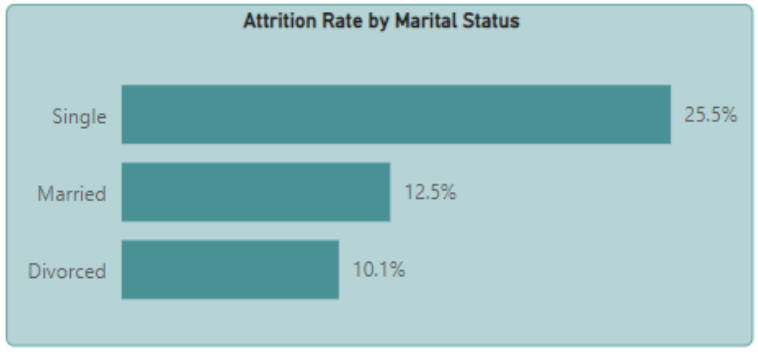


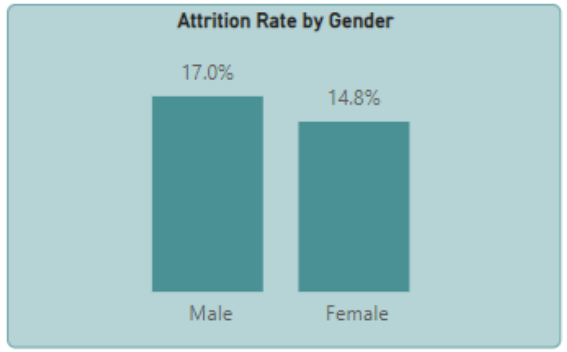


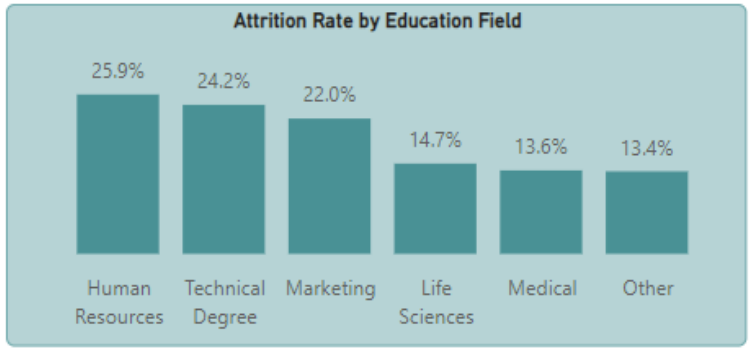






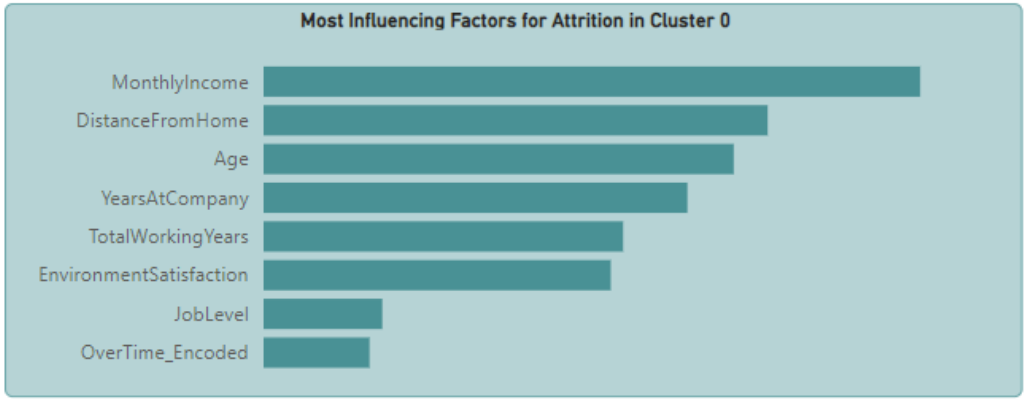


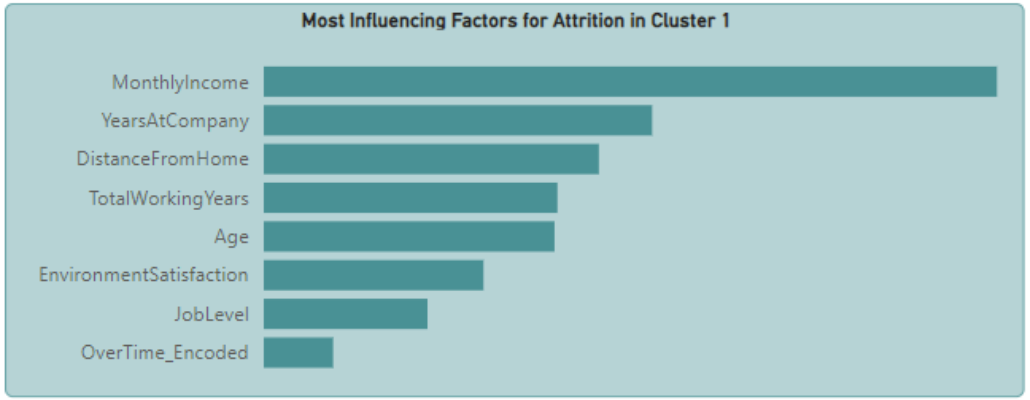




EMPLOYEE ATTRITION DASHBOARD

Basic Facts | Data Exploration | Prediction





High Risk Employees (of Attrition) in Cluster 0

Emp ID	Department	Job Level	Job Role	Age	Tenure (Y)
587	Research & Development	3	Healthcare Representative	31	1
1033	Research & Development	2	Healthcare Representative	37	1
811	Research & Development	1	Laboratory Technician	23	0
1052	Research & Development	2	Laboratory Technician	36	1
1107	Research & Development	1	Laboratory Technician	26	1
1156	Research & Development	1	Laboratory Technician	18	0
1604	Research & Development	1	Laboratory Technician	28	1
991	Research & Development	1	Research Scientist	31	3
952	Sales	1	Sales Representative	25	1
1273	Sales	1	Sales Representative	25	1

High Risk Employees (of Attrition) in Cluster 1

Employee ID	Department	Job Level	Job Role	Tenure (Y)
165	Research & Development	3	Healthcare Representative	40
582	Research & Development	3	Manager	7
825	Research & Development	5	Research Director	31
1572	Research & Development	3	Manufacturing Director	33
291	Sales	3	Sales Executive	14
307	Sales	3	Sales Executive	16
970	Sales	3	Sales Executive	14
1038	Sales	5	Manager	32
1639	Sales	3	Sales Executive	13
1968	Sales	3	Sales Executive	2