# IBM HR Analytics Employee Attrition & Performance

## 1. Introduction

### a) Problem Statement

Employee attrition is a major concern for many organizations. High employee turnover rates can be costly, affecting productivity, morale, and overall business performance. A company's HR department needs to analyze its employee data to gain insights into the factors that contribute to employee attrition.

a) Here are some examples of the factors affecting employee attrition (some other factors will probably be identified as the project evolves):
   i. average monthly income
   ii. relations with team members
   iii. distance from home
   iv. job role

b) For this project, optimality is measured through the number of employees leaving a company. Therefore, the goal is to determine the parameters associated with attrition that yield the minimum number of employees leaving a company--for a given set of combinations of these parameters.

### b) Background

The company's success is heavily dependent on the performance and engagement of its employees. Therefore, it is critical for every company to understand the factors that lead to employee attrition and develop effective retention strategies. Employee attrition, or the process of employees leaving a company and needing to be replaced, can have several disadvantages for businesses:

a) Cost: Hiring and training new employees can be expensive
b) Loss of productivity: New employees need time to adjust to their role and learn about the company's culture, processes, and expectations
c) Reduced morale: High turnover rates can negatively affect employee morale and engagement
d) Knowledge loss: When experienced employees leave, they take their knowledge, skills, and institutional memory with them
e) Disruption to teamwork: When a team member leaves, it can disrupt the dynamic and workflow of the team
f) Damage to reputation: High turnover rates can damage a company's reputation as a desirable employer

## 2. Datasets

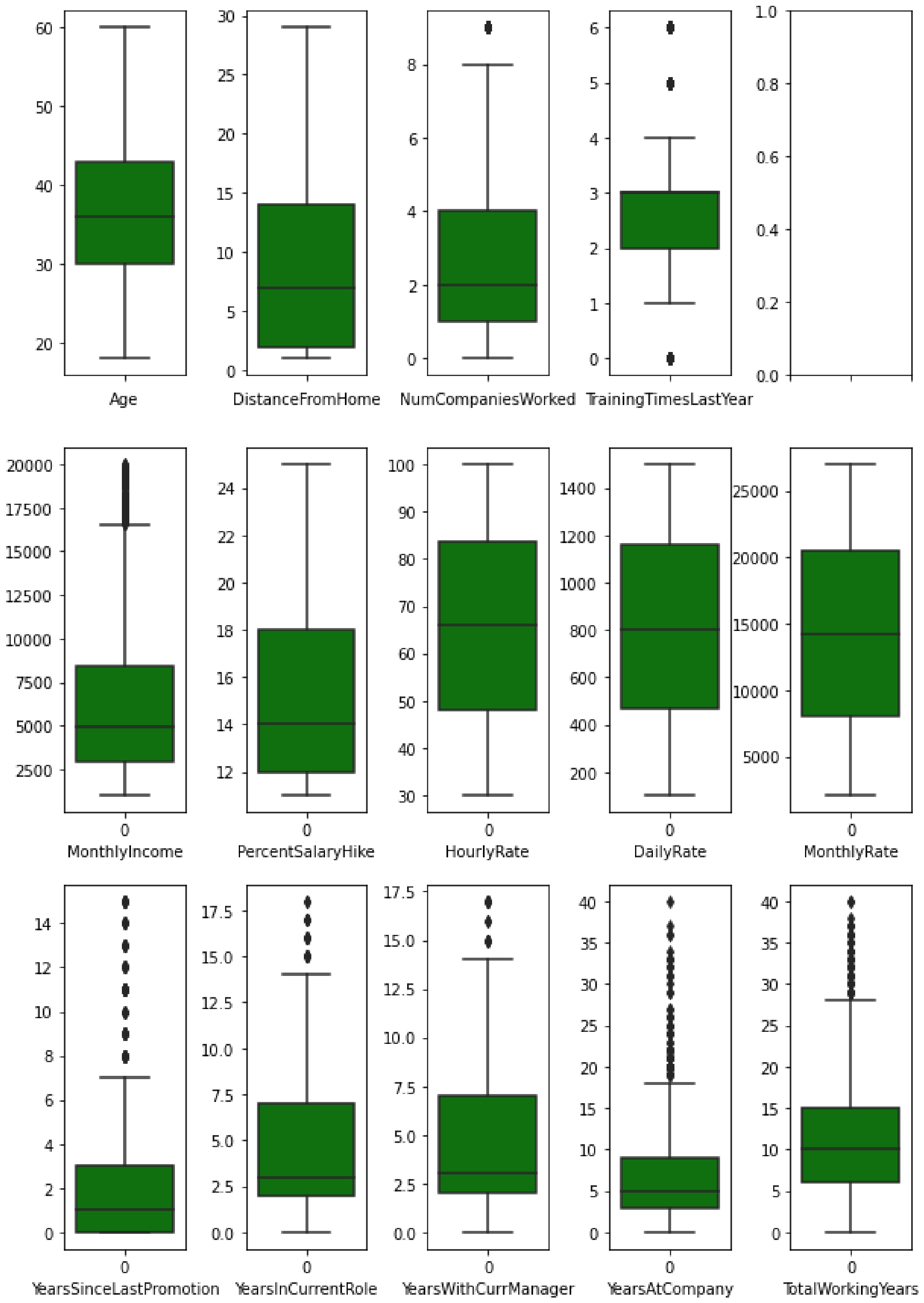
Below is the data source for this project which includes IBM's fictional HR data set created by their data scientists. This dataset contains information on employee demographics, job history, performance metrics, salary, and benefits.

**IBM HR Analytics Employee Attrition & Performance | Kaggle**

The Attrition dataset had 1470 observations with 35 variables. Out of the 35 variables, there exists one target variable Attrition with possible outcomes Yes and No. The other 34 variables are independent variables but one, that was, Employee Number which denotes the employee number or the identification number.

## 3. Data Cleaning and Data Wrangling

1. **Checking for duplicates**: Checked for duplicate rows from the initial dataset. Performed this using column 'EmployeeNumber' which is a unique number. There were no duplicates in this dataset.
2. **Checking for missing values**: The dataset has no missing values. Hence, no further treatment is required pertaining to the missing values.
3. **Feature Selection**: From modeling perspective, following 4 irrelevant columns were removed:
    a. EmployeeCount: This is just a count of employee and the value it takes is always 1.
    b. EmployeeNumber: This nominal variable is not used in the analysis as it does not provide any input to the model building process.
    c. Over18: This variable describes if an employee is over 18 years of age. It takes the value 'Yes' in all cases.
    d. StandardHour: The standard number of hours an employee works in a week. Its constant value is 80.
4. **Checking Outliers:** Created box plots to check for the outliers for all the numeric columns

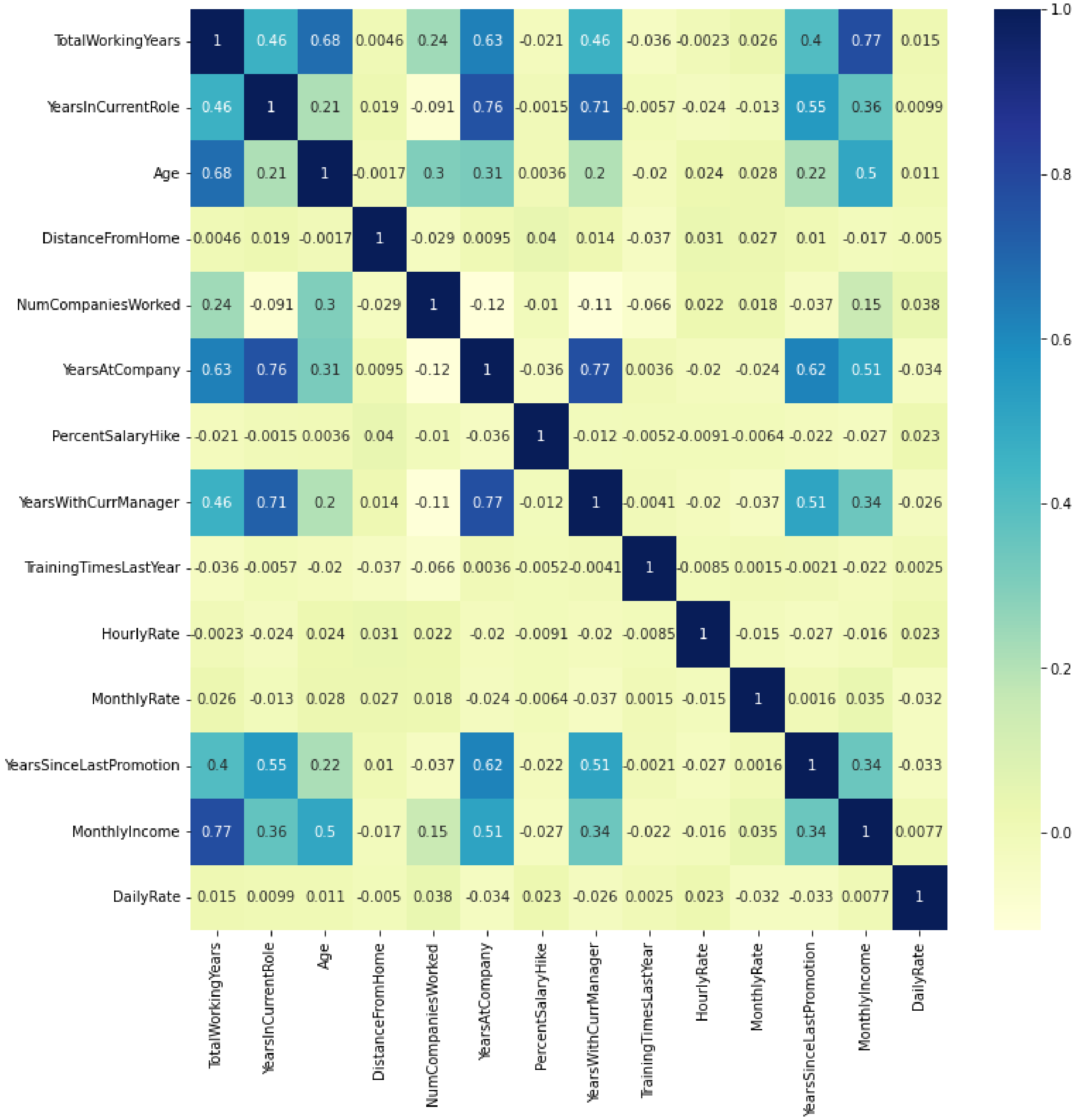

Age
60
50
40
30
20
DistanceFromHome
30
25
20
15
10
5
0
NumCompaniesWorked
8
6
4
2
0
TrainingTimesLastYear
6
5
4
3
2
1
0
1.0
0.8
0.6
0.4
0.2
0.0
MonthlyIncome
20000
17500
15000
12500
10000
7500
5000
2500
PercentSalaryHike
24
22
20
18
16
14
12
HourlyRate
100
90
80
70
60
50
40
30
DailyRate
1400
1200
1000
800
600
400
200
MonthlyRate
25000
20000
15000
10000
5000
YearsSinceLastPromotion
14
12
10
8
6
4
2
0
YearsInCurrentRole
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
YearsWithCurrManager
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
YearsAtCompany
40
35
30
25
20
15
10
5
0
TotalWorkingYears
40
35
30
25
20
15
10
5
0
0

**Observations:**

1. Age, DistanceFromHome, HourlyRate, MonthlyRate, DailyRate, PercentSalaryHike tend not to have any outliers.
2. NumCompaniesWorked, TrainingTimesLastYear, YearsWithCurrManager, YearsInCurrentRole have a moderate number of outliers.
3. MonthlyIncome, TotalWorkingYears, YearsAtCompany, YearsSinceLastPromotion have large number of outliers.

One way to counter this problem is by scaling the variables to reduce its effect on the model.

5. **Checking for multicollinearity:** One final step before moving further is to check for multi collinearity. Plotted a correlation matrix for this purpose. Below heatmap shows correlation between independent numeric variables.



**<u>Observations:</u>**

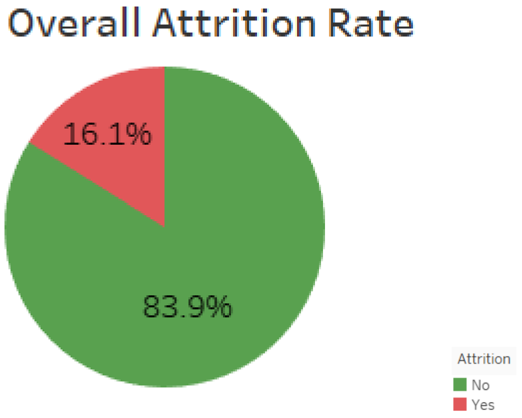
1. TotalWorkingYears: 0.77 correlation with MonthlyIncome.
2. YearsAtCompany: 0.77 with YearsWithCurrManager and 0.76 correlation with YearsInCurrentRole

No variables were removed as none of the variables have correlation more than 0.8.

6. **Preprocessing**:
    a. **Label encoding**: Encoded categorical variables using One-hot encoding to avoid multi-collinearity.
    b. **Standardization**: Standardized numerical data of both training and test dataset fitting scaler based on training data.

## 4. Exploratory Data Analysis and Initial Findings

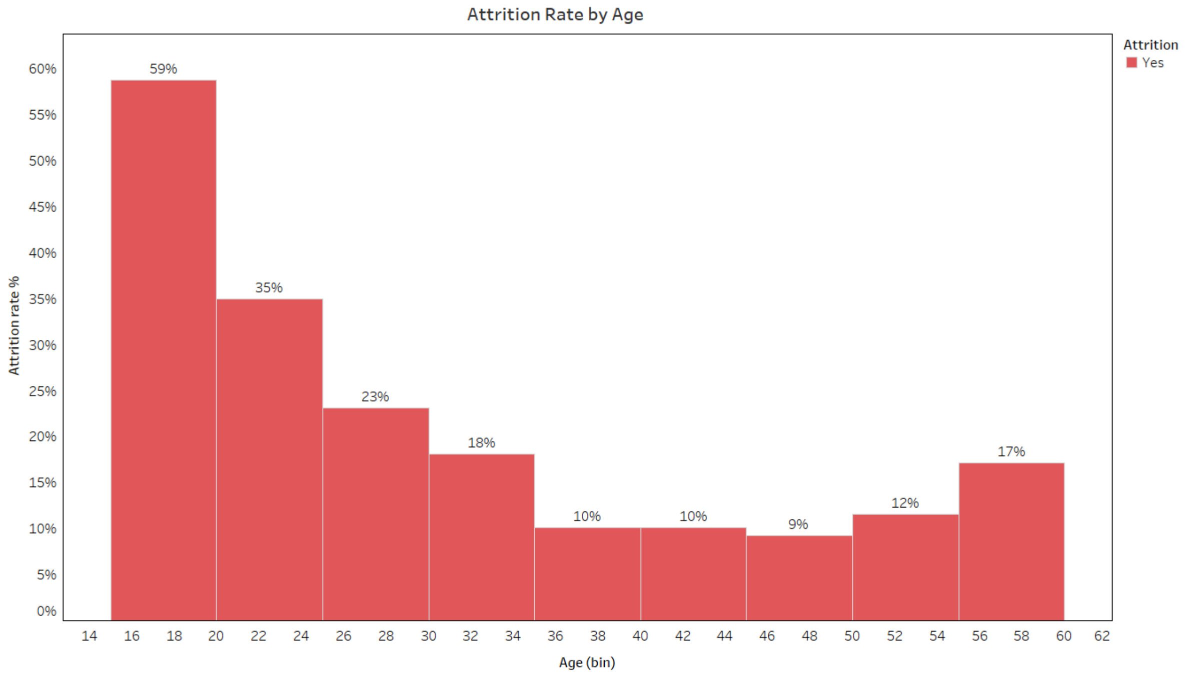
Let's explore important questions such as 'show me a breakdown of distance from home by job role and attrition' or 'compare average monthly income by education and attrition'.



The graph displays the distribution of the target variable, 'Attrition.' Out of the total 1470 observations, 16.1% (167 employees) are labeled as 'No,' indicating they did not leave, while 83.9% (1233 employees) are labeled as 'Yes,' indicating they did leave. The attrition rate is calculated by dividing the count of employees who left within a specific category by the total number of employees in that category.

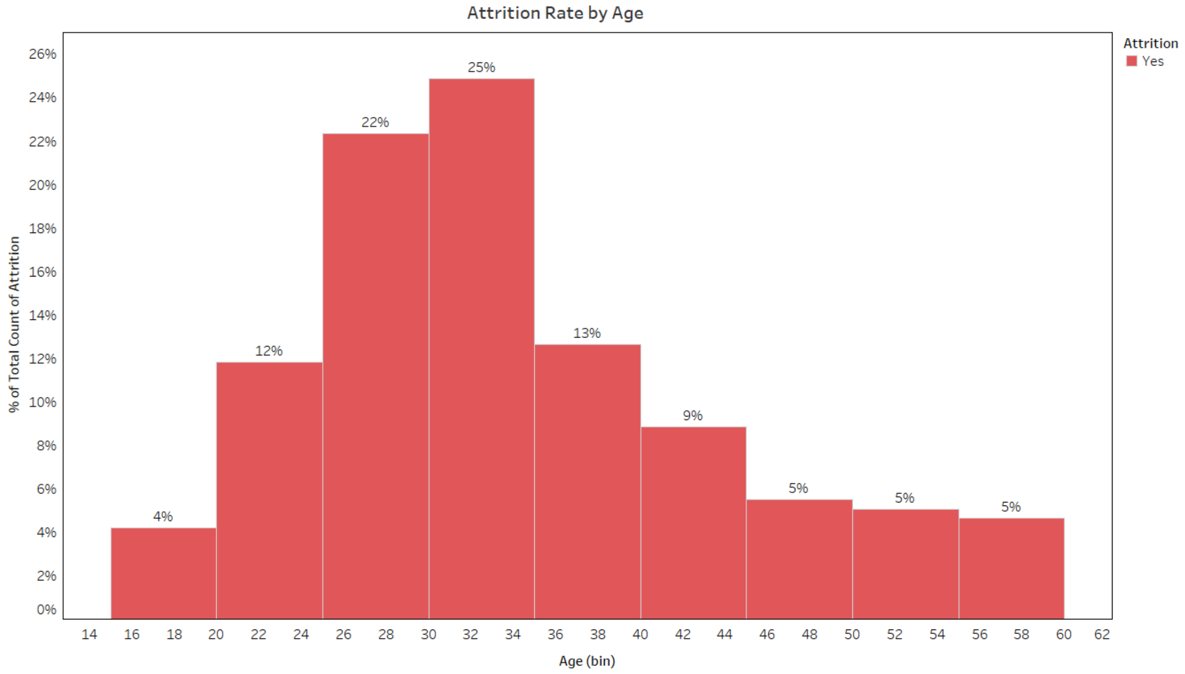
1. **Personal Factors**: First, let's examine the personal factors that contribute to employee attrition.

a. Age: Attrition is highest among employees under 20 years old, decreases with age until 50, and slightly rises for employees in their 50s.



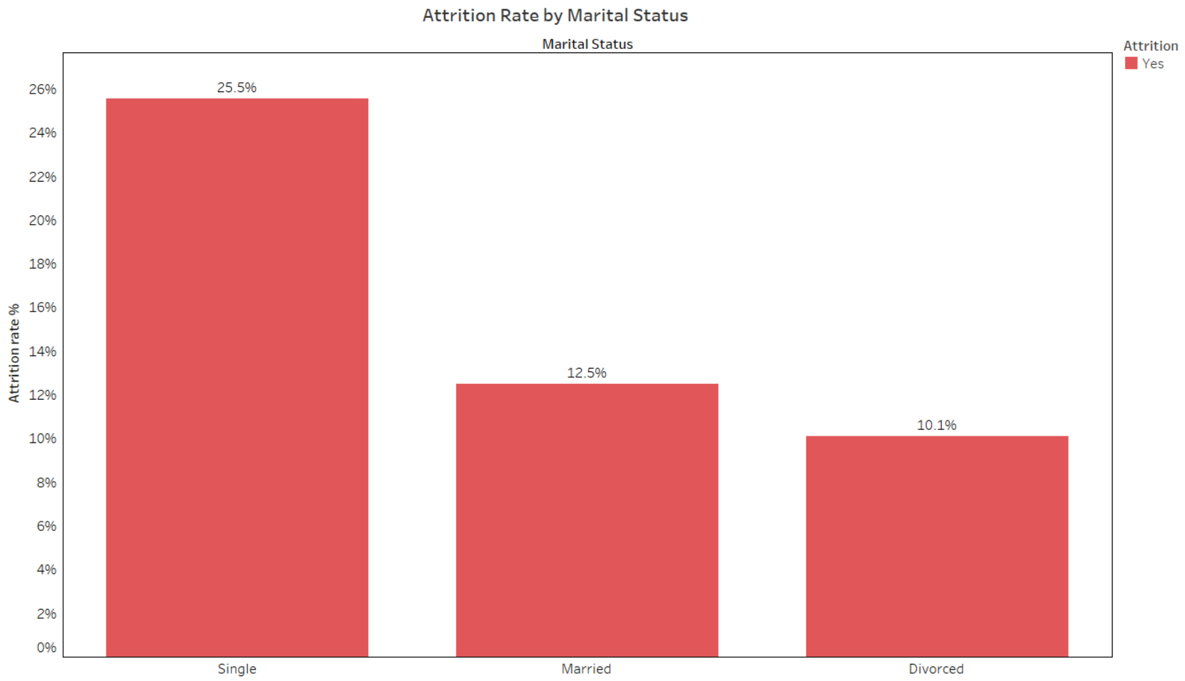
The trend of % of Total Count of Attrition for Age (bin) broken down by Attrition. Color shows details about Attrition. The marks are labeled by % of Total Count of Attrition and count of Attrition. The view is filtered on Attrition, which keeps No and Yes.

However, in terms of count, employees within the age range of 31-35 have the highest attrition rate. It is noteworthy that the company also has the highest number of employees within this age range.

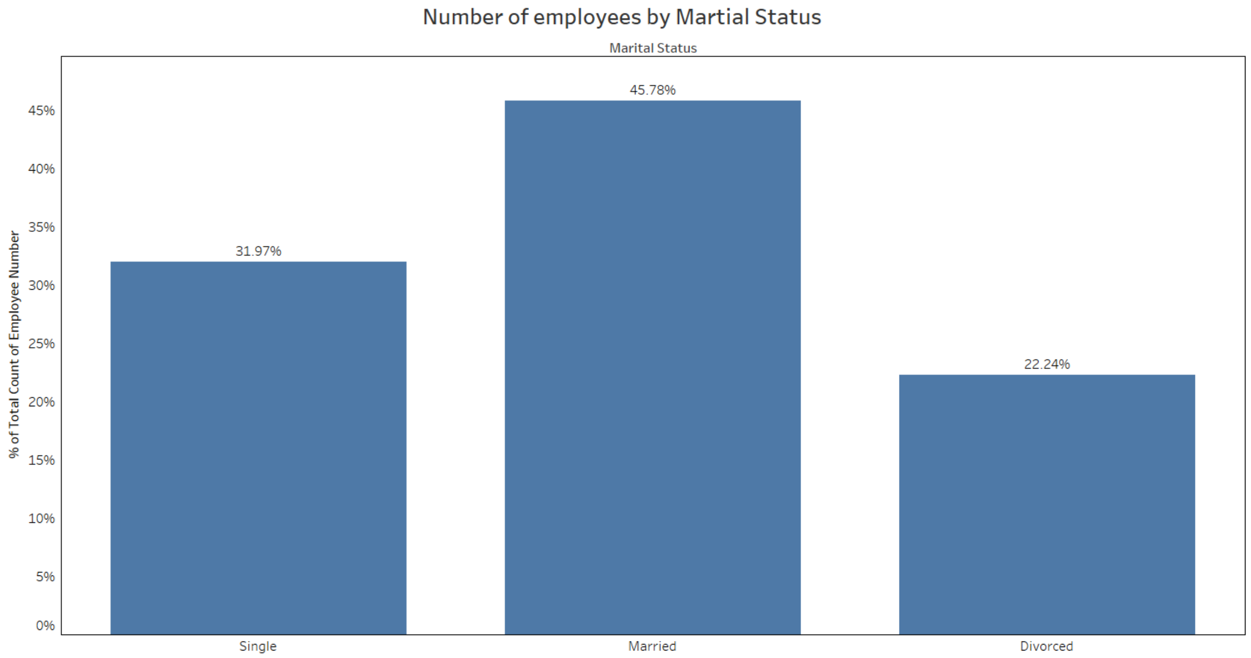


The trend of % of Total Count of Attrition for Age (bin) broken down by Attrition. Color shows details about Attrition. The marks are labeled by % of Total Count of Attrition and count of Attrition. The view is filtered on Attrition, which keeps No and Yes.

b. Marital Status: Single employees exhibit a higher attrition rate compared to married and divorced employees, but they do not comprise the largest proportion of employees in the company. The highest number of employees in the company is represented by married employees, accounting for 45.8%.

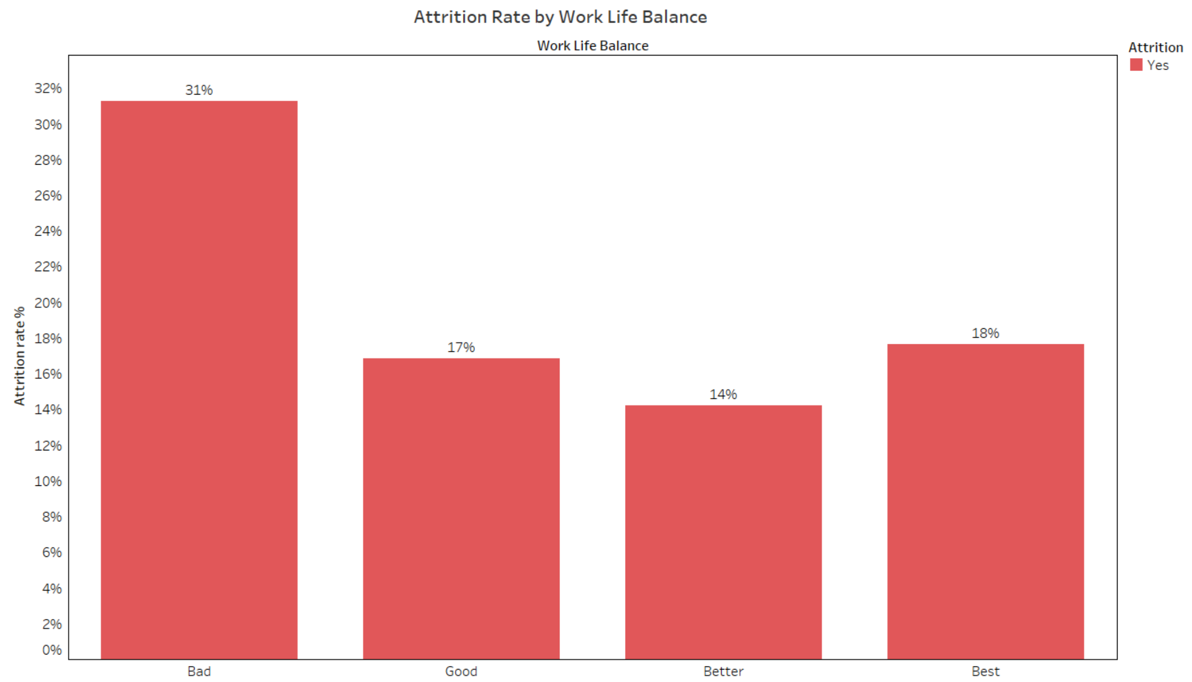

Attrition Rate by Marital Status

% of Total Count of Attrition for each Marital Status broken down by Attrition. Color shows details about Attrition. The marks are labeled by % of Total Count of Attrition and count of Attrition. The view is filtered on Attrition, which keeps No and Yes.


Number of employees by Martial Status

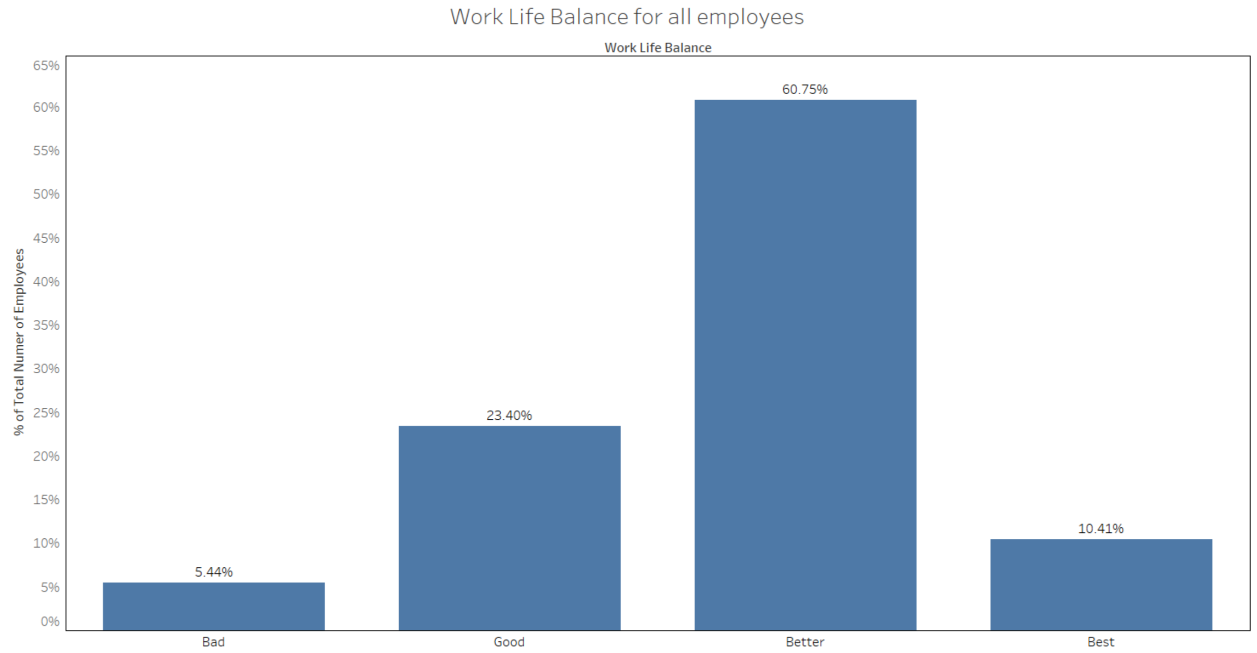
% of Total Count of Employee Number for each Marital Status. The marks are labeled by % of Total Count of Employee Number.

c. Work life balance: 31% of employees who left the company felt that their work-life balance was in a bad condition, while only 5.4% of the total number of employees felt the same. This indicates that employees with a poor work-life balance are more likely to leave the company.



Attrition Rate by Work Life Balance

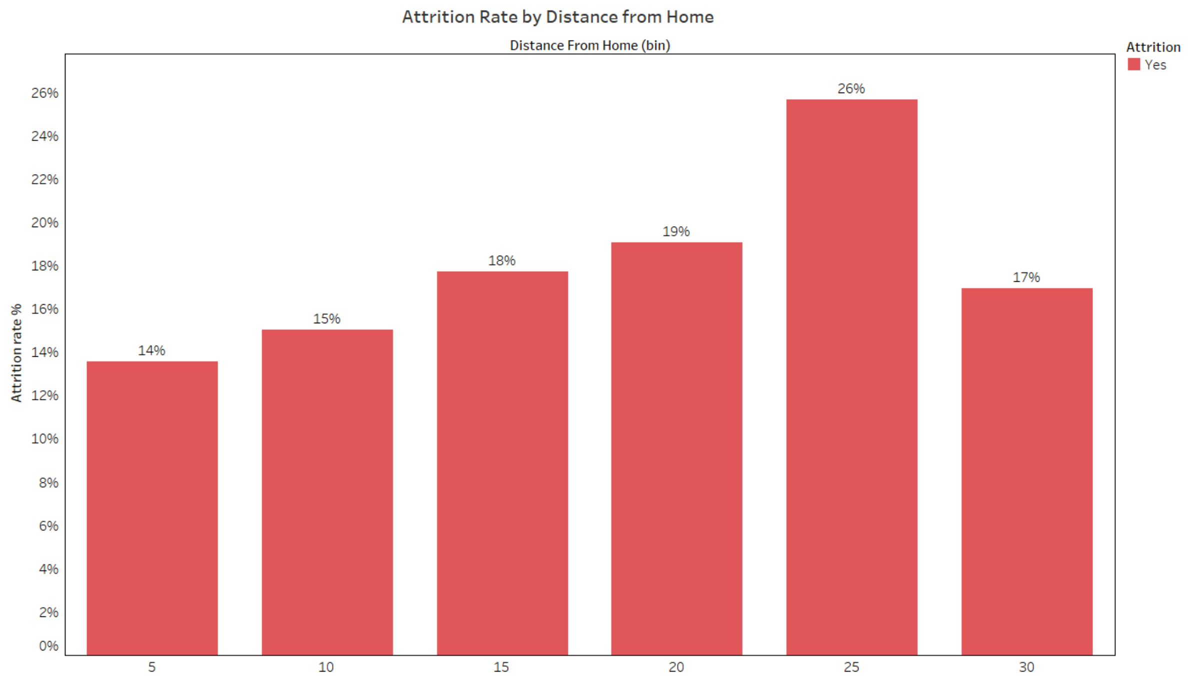
% of Total Count of Attrition for each Work Life Balance broken down by Attrition. Color shows details about Attrition. The marks are labeled by % of Total Count of Attrition and count of Attrition. The view is filtered on Attrition, which keeps No and Yes.



Work Life Balance for all employees

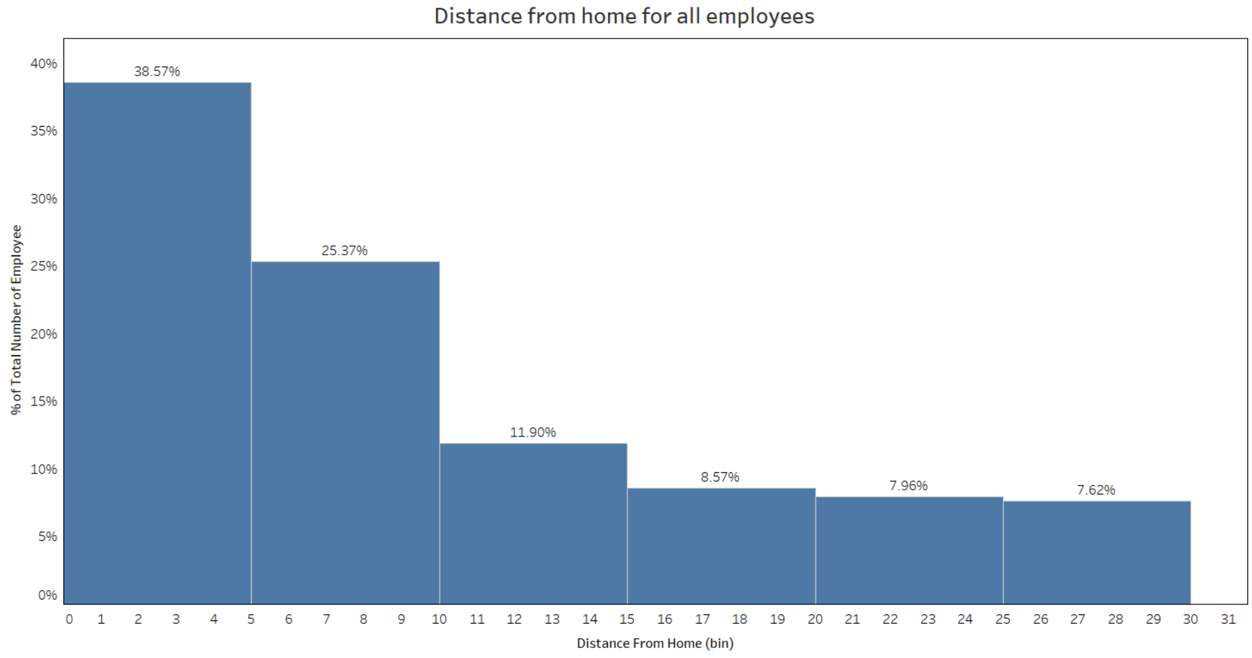
% of Total Count of Employee Number for each Work Life Balance. The marks are labeled by % of Total Count of Employee Number.

d. Distance From Home: The attrition rate exhibits a slight increase with distance from home up to 25 units, followed by a subsequent drop.



% of Total Count of Distance From Home for each Distance From Home (bin) broken down by Attrition. Color shows details about Attrition. The marks are labeled by % of Total Count of Distance From Home and count of Attrition. The view is filtered on Attrition, which keeps No and Yes.

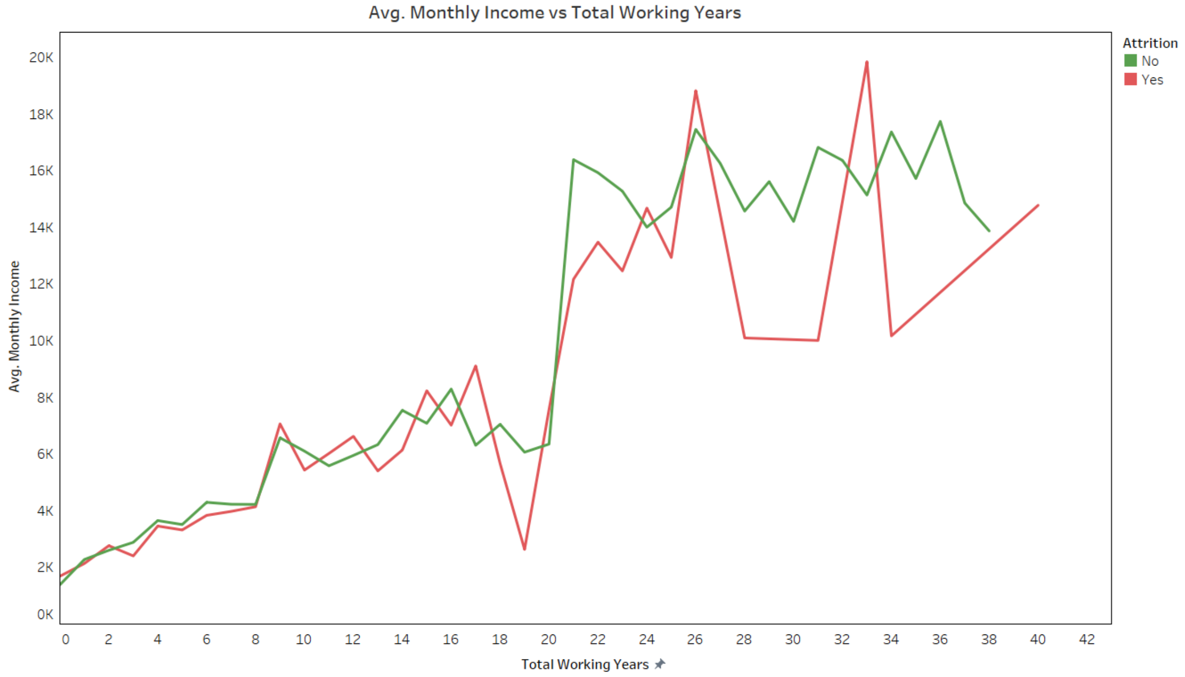
Below graph shows that the majority of employees work within a 10-mile radius from their homes, but form the above graph it can be inferred that as the distance increases, the percentage of attrition also increases.



The trend of % of Total Count of Employee Number for Distance From Home Lower Bound(bin). The marks are labeled by % of Total Count of Employee Number.

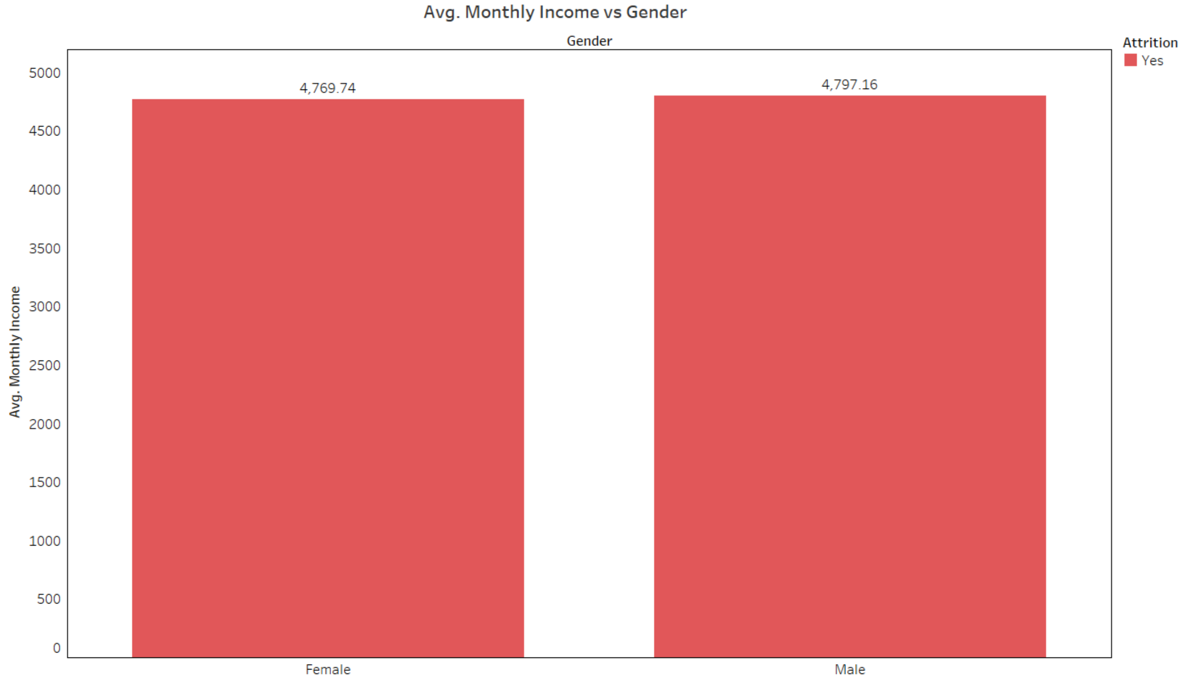
2. **Compensation**:
   a. <u>By monthly income and total working years</u>: Average monthly income tends to increase with total working years, with a higher rate for employees without attrition compared to those with

attrition.



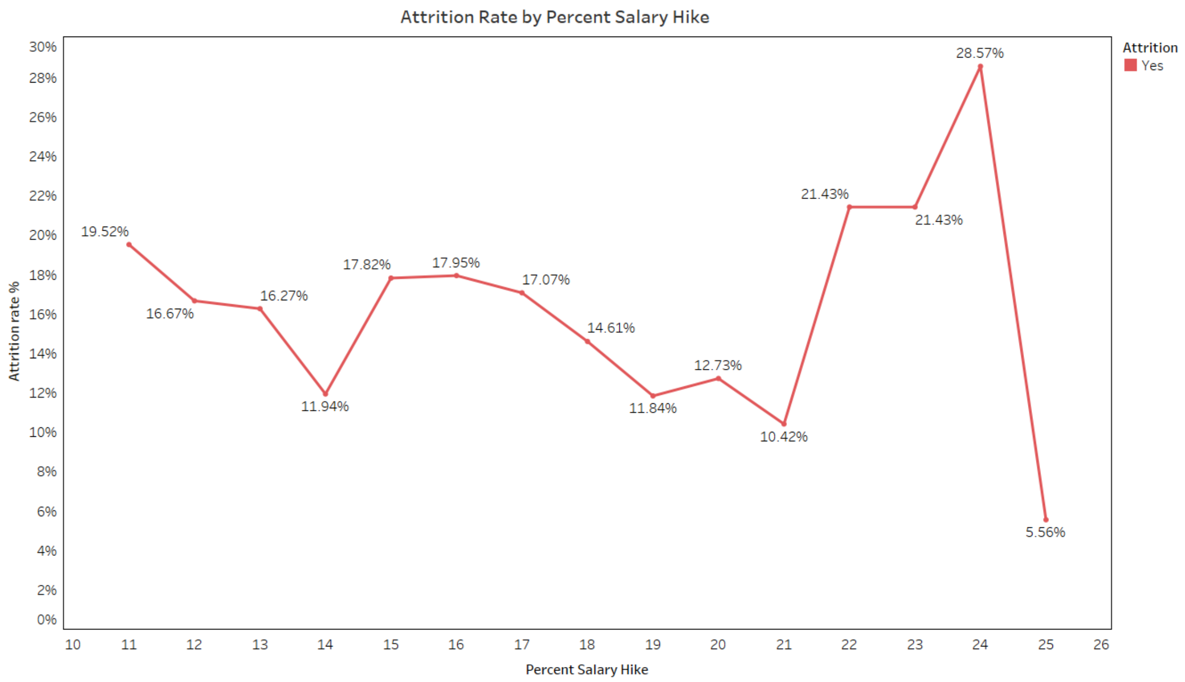
The trend of average of Monthly Income for Total Working Years. Color shows details about Attrition. The view is filtered on Attrition, which keeps No and Yes.

b. By monthly income and gender: There is no gender-based salary difference among attrition employees, as both male and female employees have similar average monthly income. Gender does not appear to have a significant impact on determining attrition.



Average of Monthly Income for each Gender broken down by Attrition. Color shows details about Attrition. The marks are labeled by average of Monthly Income and count of Attrition. The view is filtered on Attrition, which keeps Yes.

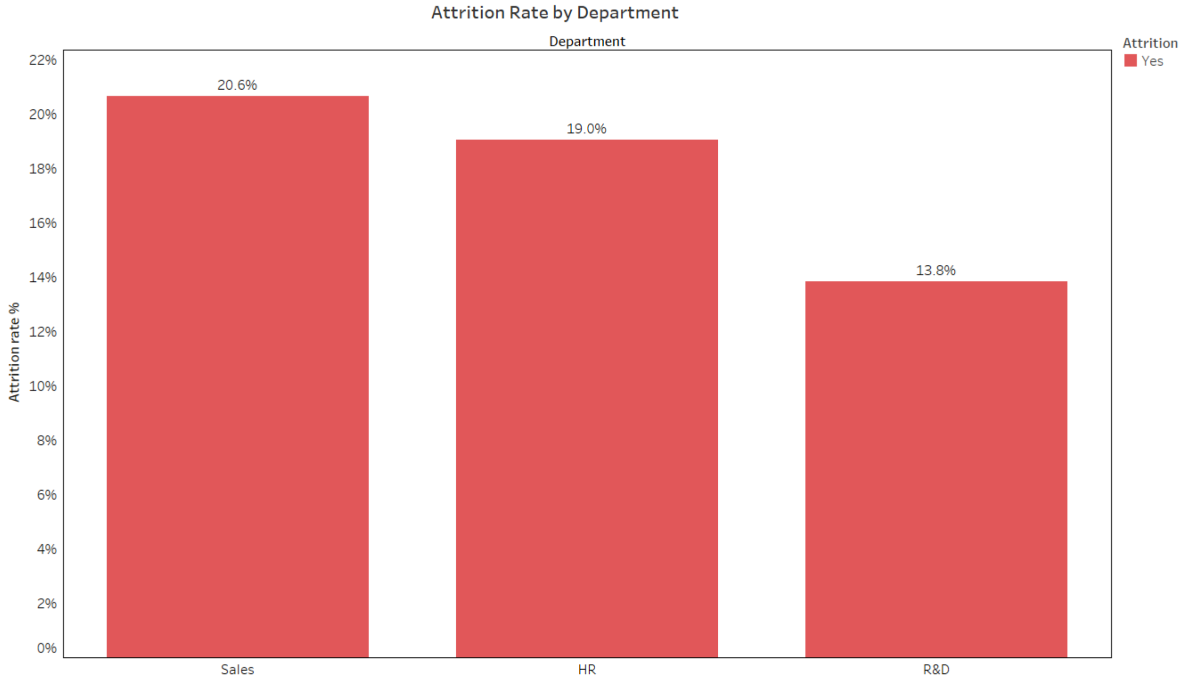
c. By percent salary hike: Generally, as the percent salary hike increases, the percentage of employees who left the company decreases with few exceptions percent salary hikes (such as 22%, 23%, and 24%).



The trend of % of Total Count of Employee Number for Percent Salary Hike broken down by Attrition. Color shows details about Attrition. The marks are labeled by % of Total Count of Employee Number. The view is filtered on Attrition, which keeps No and Yes.

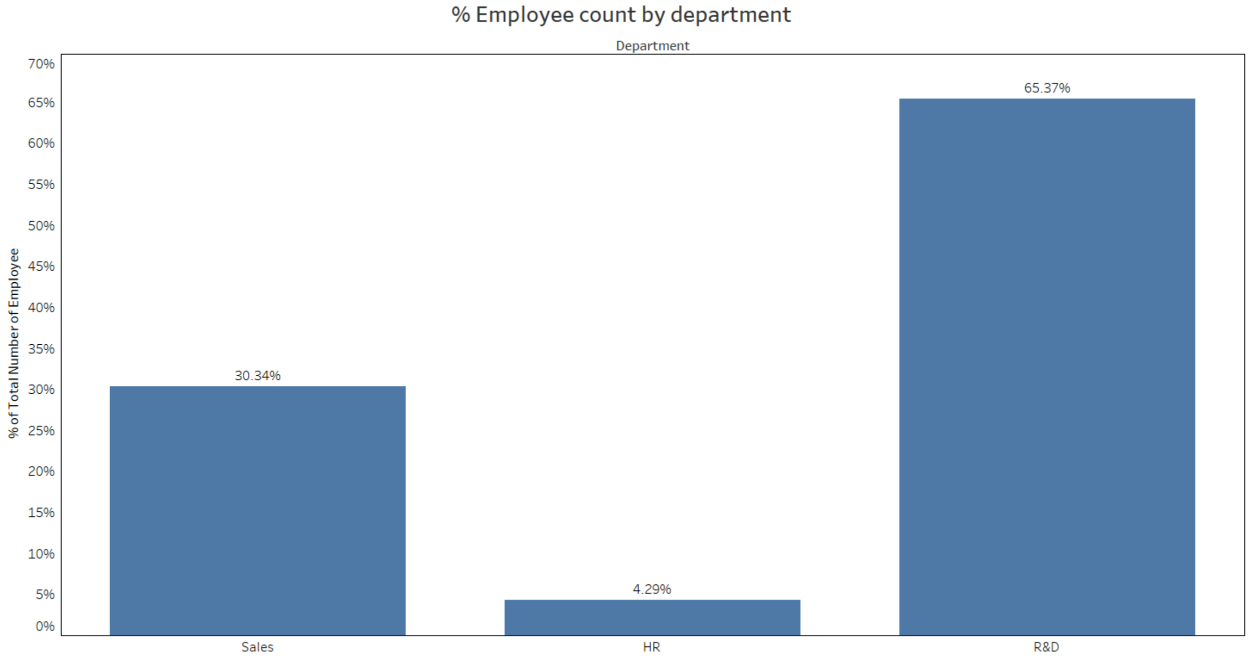
3. **Job level:**

   a. By department: The attrition rate varies across departments and job levels. The Sales department has the highest attrition rate, with 20.6% of employees leaving. The HR department has the second-highest attrition rate at 19%, followed by the R&D department with a relatively lower attrition rate of 13.8%.



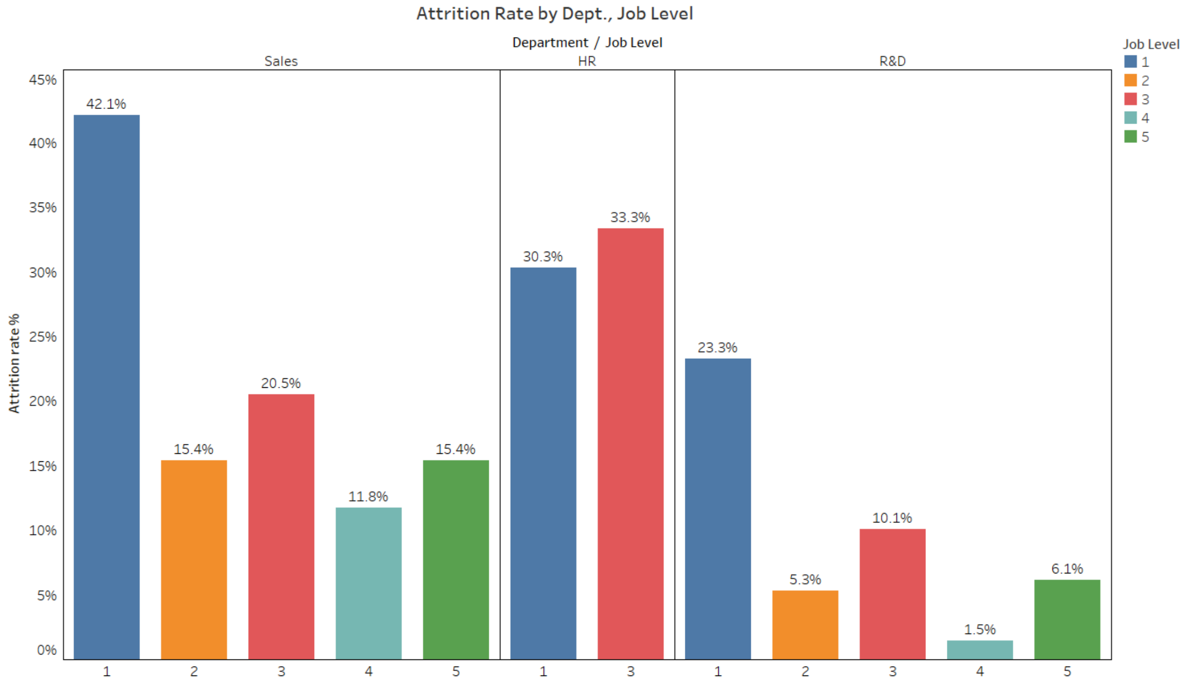
% of Total Count of Attrition for each Department broken down by Attrition. Color shows details about Attrition. The marks are labeled by % of Total Count of Attrition. The view is filtered on Attrition, which keeps No and Yes.

The total number of employees is highest in Research & Development (R&D) department with 65.37% of total employees, but its attrition rate is lowest among all the 3 departments.



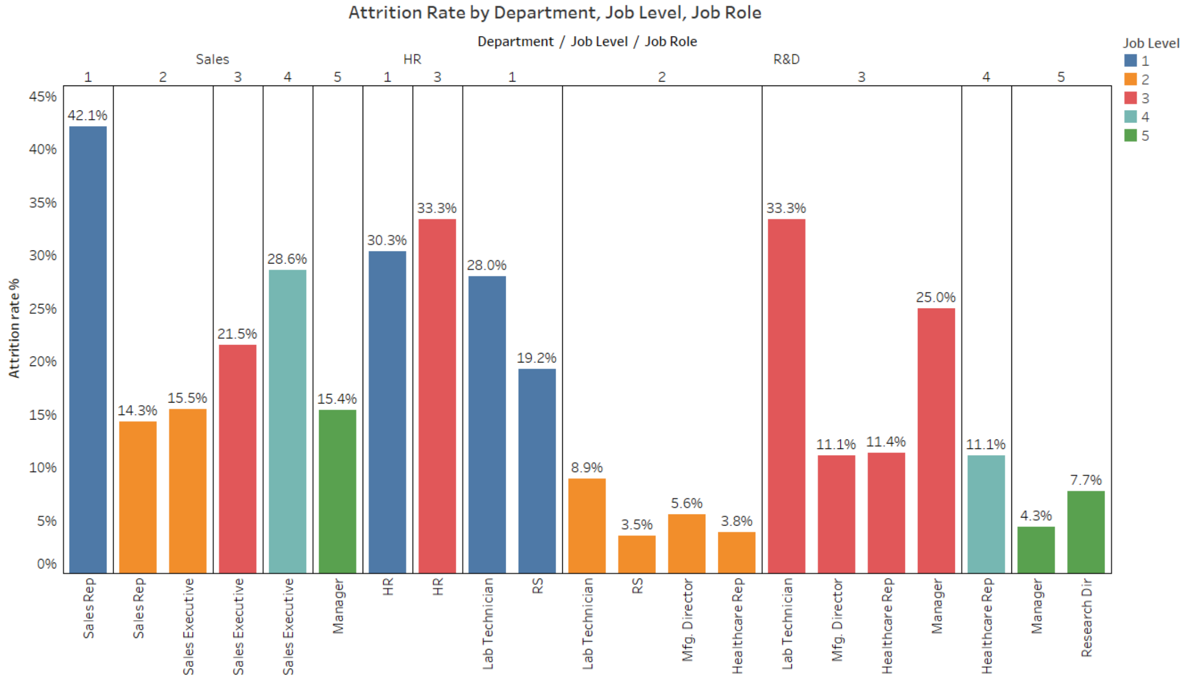
% of Total Count of Employee Number for each Department. The marks are labeled by % of Total Count of Employee Number.

b. By department and Job level: The numerical scale in the below graph ranges from 1, indicating the most junior positions, to 5, representing the most senior positions. It can be observed from the below graph that junior most position has the highest attrition rate in Sales and R&D departments.



% of Total Count of Attrition for each Job Level broken down by Department vs. Attrition. Color shows details about Job Level. The marks are labeled by % of Total Count of Attrition. The view is filtered on Attrition, which keeps No and Yes.

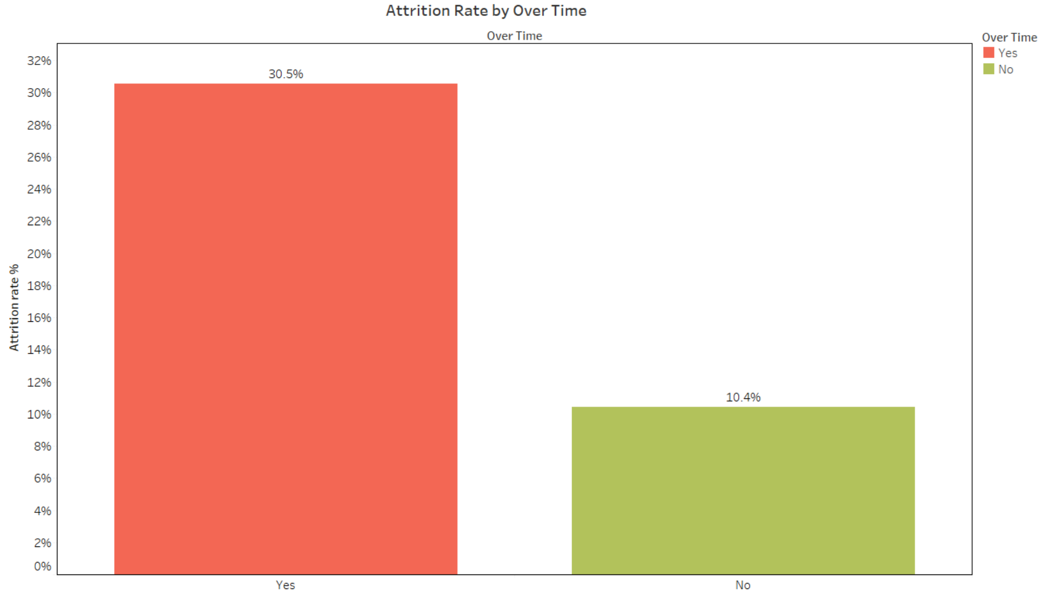
c. By department, job level and job role: Junior positions such as Sales Representative, Lab Technician, and Research Scientist contribute to high attrition rate of company.



% of Total Count of Attrition for each Job Role broken down by Department and Job Level vs. Attrition. Color shows details about Job Level. The marks are labeled by % of Total Count of Attrition. The view is filtered on Attrition, which keeps No and Yes.

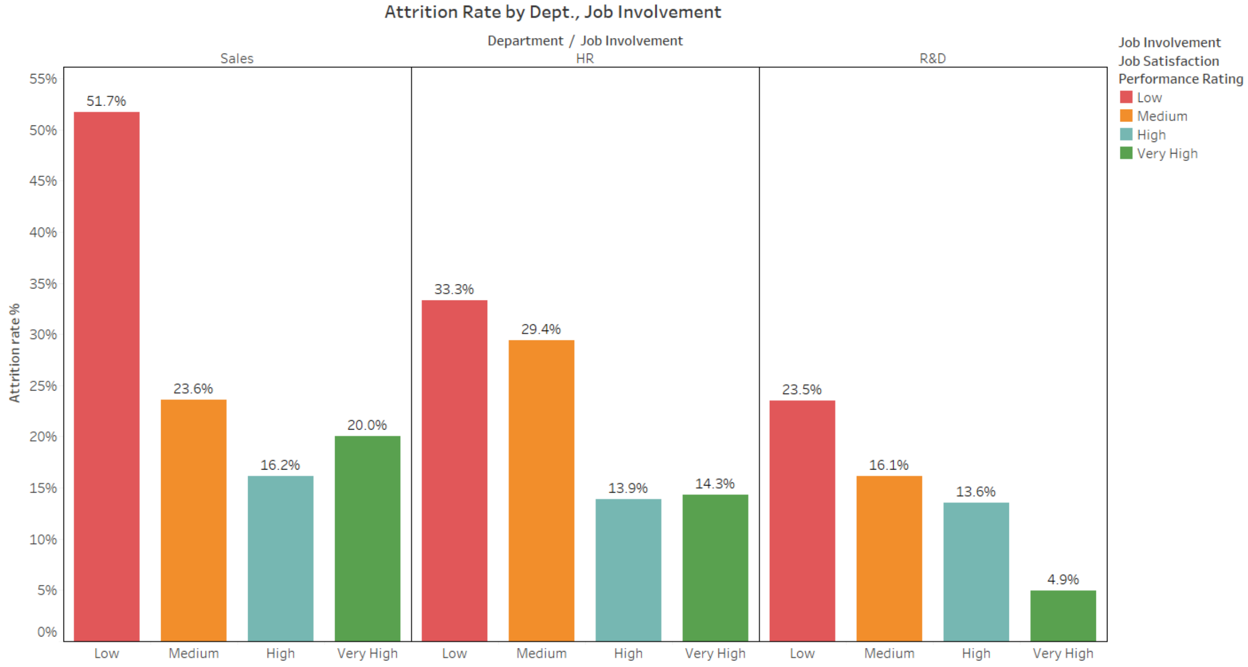
4. **Work Environment**:

a. By Overtime: Out of the total number of employees who left the company, 10.44% of employees did not work overtime, and 30.53% of employees did work overtime. The percentage of attrition is significantly higher for employees who work overtime compared to those who do not work overtime. This suggests that there may be a correlation between working overtime and employee attrition.

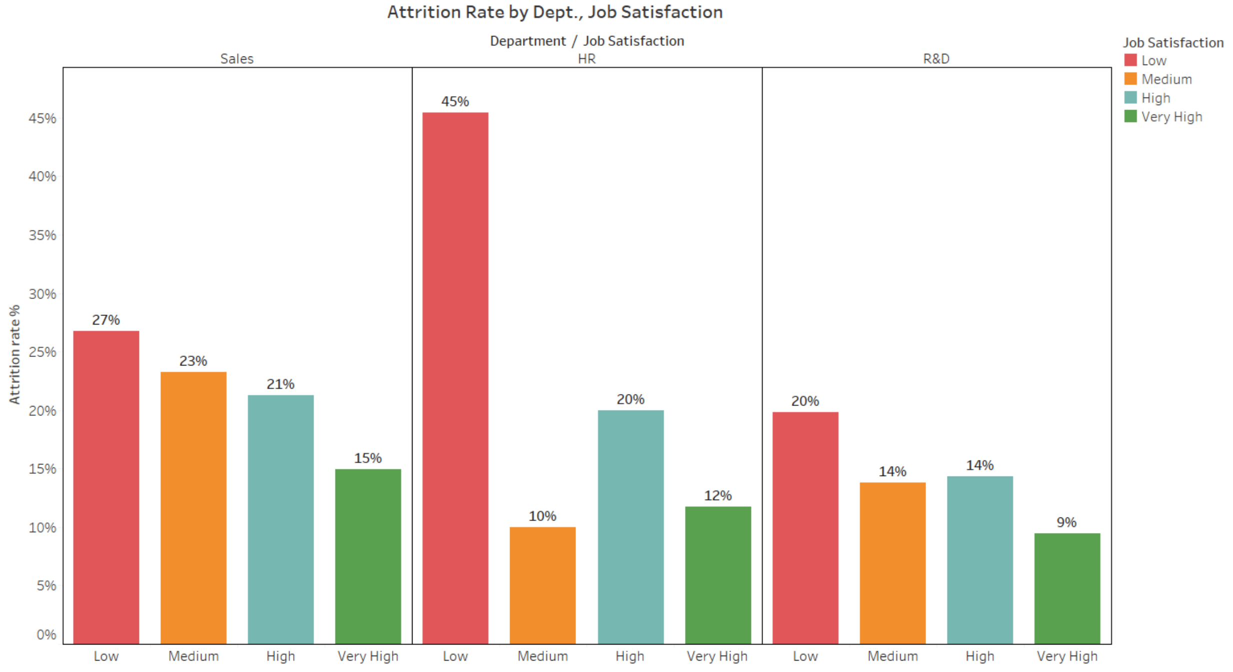


% of Total Count of Attrition for each Over Time broken down by Attrition. Color shows details about Over Time. The marks are labeled by % of Total Count of Attrition and count of Attrition. The view is filtered on Attrition, which keeps No and Yes.

b. By job involvement: Low job involvement is consistently associated with higher attrition rates across all three departments. Additionally, as the level of job involvement increases, the attrition rate tends to decrease.

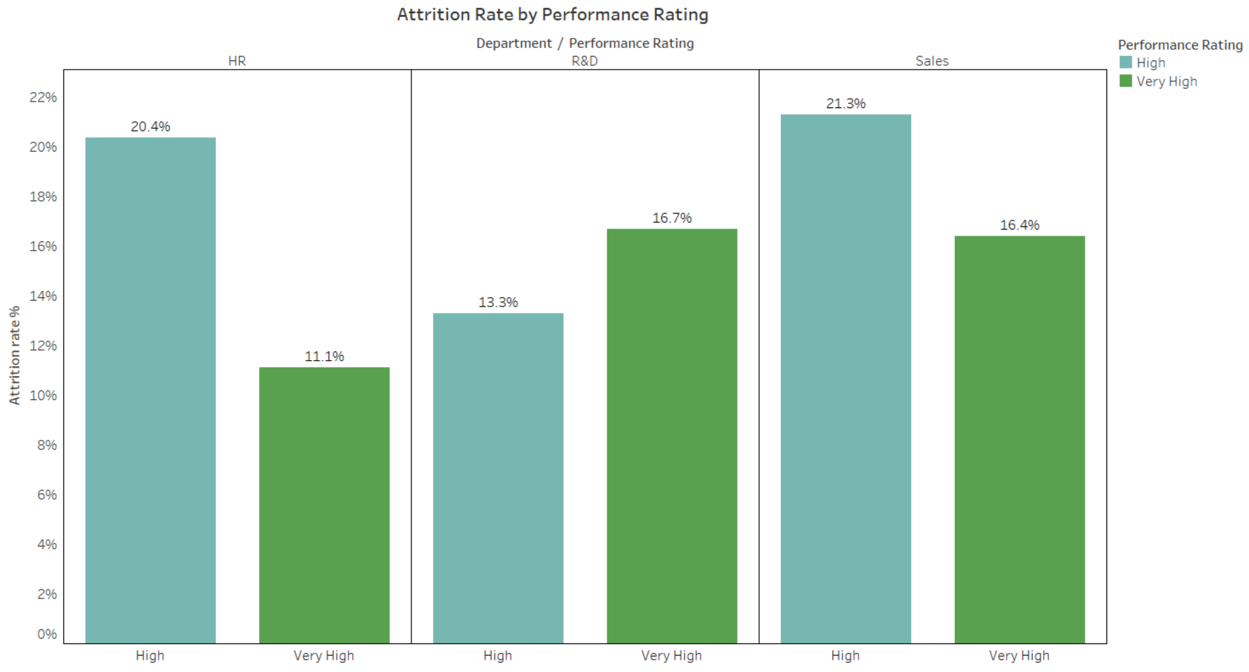


c. By job satisfaction: Low job satisfaction is consistently associated with higher attrition rates across all three departments. Additionally, as the level of job satisfaction increases, the attrition rate tends to decrease. HR department has the highest attrition rate with 45% among employees with low job satisfaction. In the R&D department, the attrition rate is relatively lower, ranging from 9% to 20%, across different levels of job satisfaction.



% of Total Count of Attrition for each Job Satisfaction broken down by Department vs. Attrition. Color shows details about Job Satisfaction. The marks are labeled by % of Total Count of Attrition. The view is filtered on Attrition, which keeps No and Yes.

d. <u>By performance rating</u>: Performance rating does not impact attrition rate as attrition employees had High and Very High rating. No employee has been assigned low or medium rating.



% of Total Count of Attrition for each Performance Rating broken down by Department vs. Attrition. Color shows details about Performance Rating. The marks are labeled by % of Total Count of Attrition. The view is filtered on Attrition, which keeps No and Yes.

## 5. Model Building

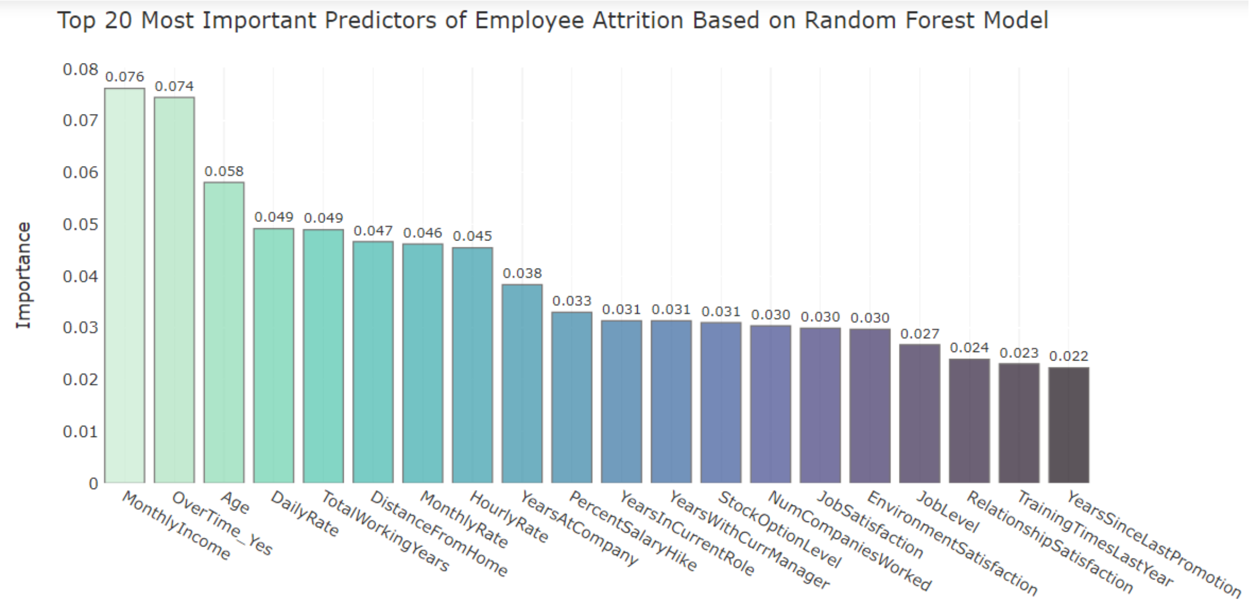
First, I transformed the categorical variables into dummy variables. I also split the data into train and tests sets with a test size of 30%.

These are the 5 models:

1. Logistic Regression – Baseline and a less complex model
2. Random Forest – Random Forest can handle imbalanced datasets where the number of instances in different classes is not balanced. It automatically balances the class distribution by bootstrapping the training samples and uses the voting mechanism to assign the final class.
3. Random Forest with hyperparameter tuning – Again, with the sparsity associated with the data, I thought that this would be a good fit. I tuned following hyperparameters for this model:

| Hyperparameter | Value |
|---|---|
| n_estimators | 200 to 2000 |
| Max_depth | None and 10 to 100 |
| min_samples_split | [2, 4, 6, 8, 10] |
| min_samples_leaf | 1 to 4 |
| Criterion | gini, entropy |
| max_features | auto, sqrt, log2 |

**Feature Importance:**

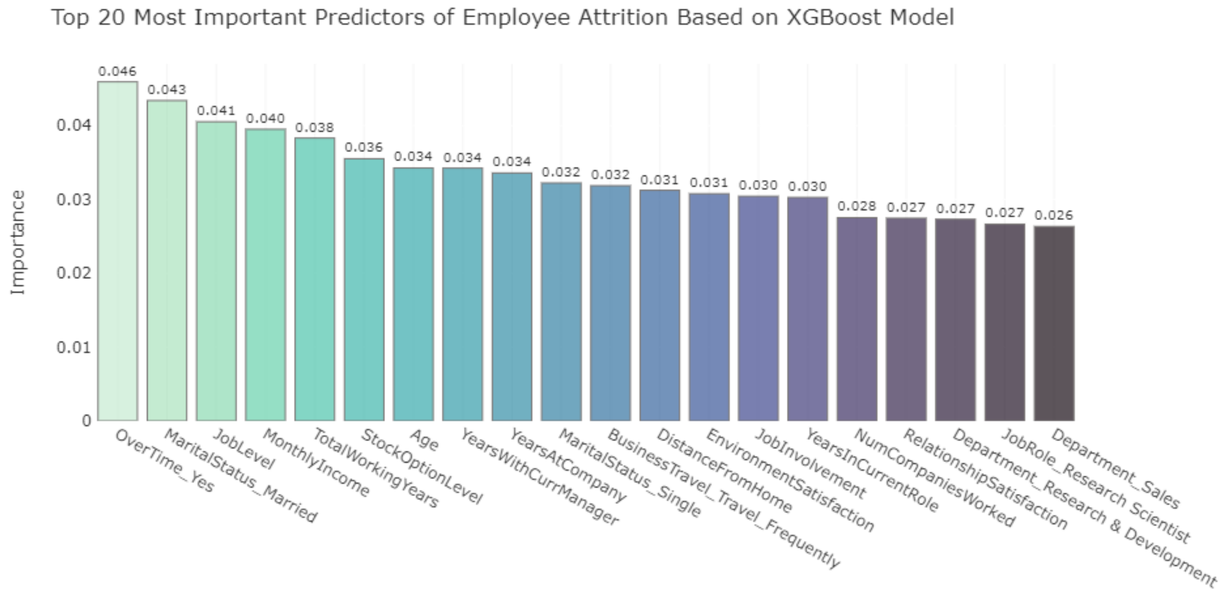


4. XGBoost – I thought this would remove any overfitting.
5. XGBoost with hyperparameter tuning - I tuned following hyperparameters for this model:

| Hyperparameter | Value |
|---|---|
| n_estimators | 100, 200, 500, 750 |
| Learning Rate | 0.01, 0.02, 0.05, 0.10, 0.25 |

| | |
|---|---|
| Max_depth | 3 to 5, 10, 12 |
| Gamma | 0.1, 0.5, 1, 1.5, 5 |
| Subsample | 0.6, 0.8, 1.0 |
| colsample_bytree | 0.6, 0.8, 1.0 |
| Min_child_weight | 1, 5, 7, 10 |

**Feature Importance:**



## 6. Model performance

The XGBoost model far outperformed the other approaches on the test and validation sets.

| **Algorithm** | **Area Under the Precision-Recall Curve (PR AUC)** | **Macro Avg Recall** | **Accuracy** |
|---|---|---|---|
| Logistic Regression | 0.62 | 0.69 | 87.07% |
| Random Forest | 0.53 | 0.57 | 85.49% |
| Random Forest with hyperparameter tuning | 0.57 | 0.56 | 85.71% |
| XGBoost | 0.56 | 0.63 | 85.49% |
| XGBoost with hyperparameter tuning | 0.62 | 0.63 | 87.30% |

I tried five different models and evaluated them using **PR AUC**. Reasons to choose this metric:

- **Precision-Recall Area Under the Curve (PR AUC**) – I opted for the PR AUC instead ROC AUC because for imbalanced datasets PR AUC provides a more accurate representation of the classifier's performance, particularly in situations where correctly identifying the positive instances is more critical than minimizing false positives.

I also used this metric as a scoring function in cross-validation.

The other important metric used was:

- First why recall? - I chose Recall because in this case, it is more important to identify all employees who will leave, even if it results in a higher number of false positive predictions. Missing employees who will attrition can lead to lost resources, productivity, and replacement costs. Recall is crucial when the cost of false negatives is high, as it ensures a higher percentage of positive instances are captured.
- Why macro avg? - I opted for the macro average instead of the weighted average because for imbalanced datasets, macro avg avoids bias towards the majority class and provides a fair evaluation of performance across all classes, including the minority class.

## 7. Future Work

1. To add more sources for attrition to make the dataset more realistic
2. To productionize this model

## 8. Business plan to reduce employee attrition

Based on the given factors, here is a suggested employee attrition plan:

1. **Manage Overtime**: Implement strategies to reduce or manage overtime for employees. This can include workload distribution, better resource planning, and promoting work-life balance to reduce the likelihood of attrition among employees who work overtime.
2. **Support Marital Relationships**: Provide support and resources to married employees to help them maintain a healthy work-life balance. This can include flexible work arrangements, employee assistance programs, and communication channels to address any concerns or challenges they may face.
3. **Career Development and Promotions:** Create opportunities for career growth and development within the organization. Offer training programs, mentoring, and clear paths for advancement to encourage employees to stay and progress within the company. Focus on providing a clear career trajectory and recognizing and rewarding employees for their achievements.
4. **Competitive Compensation:** Ensure that the company offers competitive compensation packages, including higher monthly income or wages for employees. Regularly review and benchmark salaries to ensure they are aligned with industry standards and provide financial incentives to retain valuable employees.
5. **Employee Retention Programs**: Develop employee retention programs that specifically target employees who have reached a certain threshold of total working years (e.g., after year 3). These programs can include additional benefits, recognition, and opportunities for growth to motivate and retain experienced employees.

It's important to note that employee attrition is a complex issue influenced by various factors, and the suggested plan should be tailored to the specific needs and culture of the organization. Regularly monitor attrition rates, conduct employee surveys, and gather feedback to make adjustments and continuously improve the attrition plan.