

# **PROJECT REPORT**

## **IBM HR Analytics Employee Attrition & Performance**

---

*Submitted By:*

***Rahul Satpute***

## Introduction:

Attrition, in Human Resource terminology, refers to the phenomenon of the employees leaving the company. Attrition in a company is usually measured with a metric called attrition rate, which simply measures the no of employees moving out of the company (voluntary resigning or laid off by the company).

In this project, I want to predict the attrition of the company's valuable employees, uncover the factors that lead to employee attrition and 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'.

My client is IBM human resources director. He is trying to figure out the roots of employee attrition and improve the performance of company. For that, he focuses on defining the parameters, which cause the employee attrition via proactive approach, and tries to overcome that/those with the project's outcome.

## Data Acquisition and Cleaning

### 1. Data Acquisition

I use IBM HR Analytics Employee Attrition & Performance data from Kaggle, which is created by IBM data scientists. Dataset is in the open source website and can be reached from this link. It has 1470 rows x 35 columns and contains numeric and categorical data types in columns. I loaded the dataset from this link in csv format and read it in the Jupyter notebook after importing necessary libraries.

### 2. Data Specification:

The dataset has 1470 rows and 35 columns. Rows are observations from each employee and columns are from different features, which are obtained in order to explain the employee attrition. The features data types consist of 27 integers and 8 objects. List of attributes are presented below.

**Response Variable :** Attrition (int64)

Features: Type(Int64)	Type(Object)
Age, DailyRate, DistanceFromHome, Education, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction, HourlyRate, JobInvolvement, JobLevel, JobSatisfaction, MonthlyIncome, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager	BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, Over18, OverTime,

## Data Cleaning:

I searched for missing values in every features of dataset, all features look like having 1470 non-null entries. However, missing values can be encoded in a number of different ways, such as by zeroes. For that reason, I checked both missing values and duplicate values in the dataset. Luckily, it was okay to continue to next step.

I observed 5 random sample records in the dataset to grasp the general intuition about whole picture. Besides that, I explored the statistical attributes of each features such as their mean, standard deviation, interquartile values in order to detect outliers. This research also gave me a general impression about unique and top values for each attribute in addition to their frequencies in the dataset. I made double checks on some of features in order to make sure that everything is good to go. Those results were also okay.

I inspected the useless features in order to drop in the dataset. "Over 18", "StandardHours", and "EmployeeCount" had only one unique value for each observation and that did not impact or change anything in the data. For that reason, I dropped those three useless columns.

To be able to use effectively in the further steps, I reassigned the response variable (Attrition) which had "Yes" and "No" values previously. They were assigned to 1 and 0 respectively. After that, I moved the response variable to the last column place.

The dataset has 8 object types which are 'BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'OverTime'. To be able have more memory usage and become fast, I changed object type to category type in the dataset.

Field	1	2	3	4
Education*	Below College	College	Bachelor	Master
Environment Satisfaction	Low	Medium	High	Very High
Job Involvement	Low	Medium	High	Very High
Job Satisfaction	Low	Medium	High	Very High
Performance Rating	Low	Good	Excellent	Outstanding
Relationship Satisfaction	Low	Medium	High	Very High
Work Life Balance	Bad	Good	Better	Best

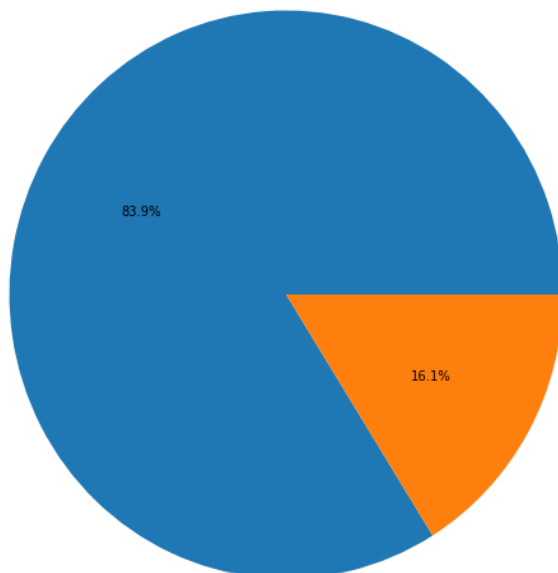
# Exploratory Data Analysis

## Introduction:

We have 32 features consisting of both categorical as well as the numerical features. Response variable is 'Attrition' of the employees, which can be 1 and 0 (representing 'Yes' and 'No' respectively). This is what we will predict.

Now, I will try to analyse visually the trends in how and why employees are quitting their jobs. For that, I will deep dive into the details about features and their relationships between each other.

## Target Variable ('Attrition'):



In the company, there are 1470 employees.

237 employees who compose 16% of the total number of employees left the company for some reasons.

Besides that, 1233 employees are currently continuing to work in the same company.

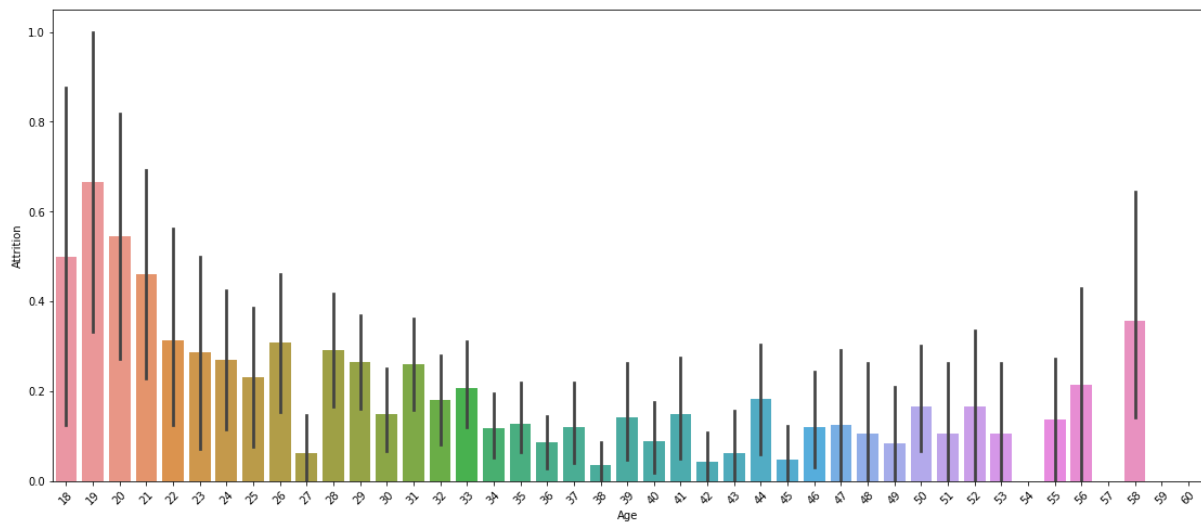
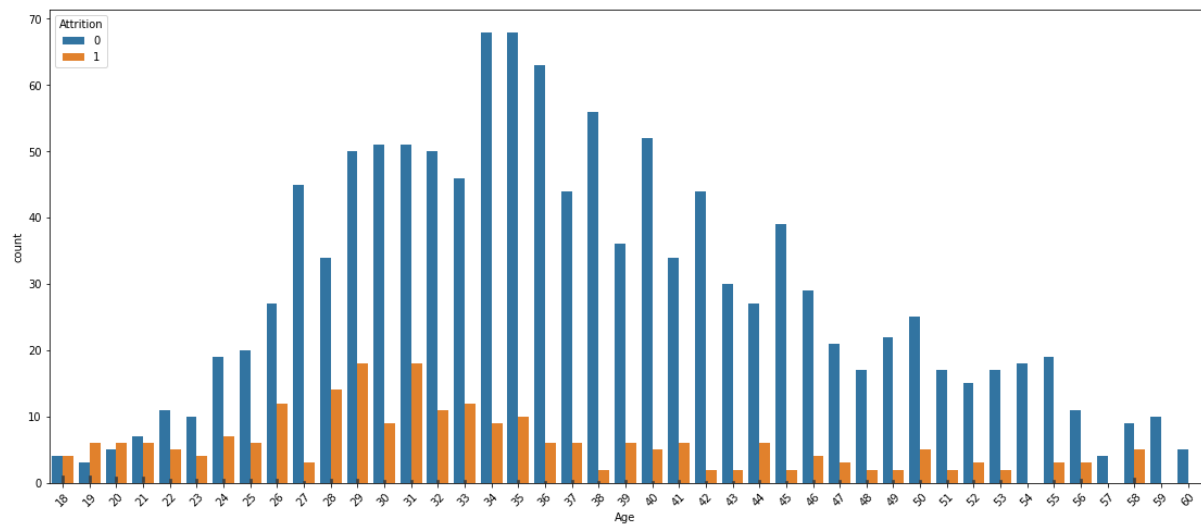
## Features:

### i. Age:

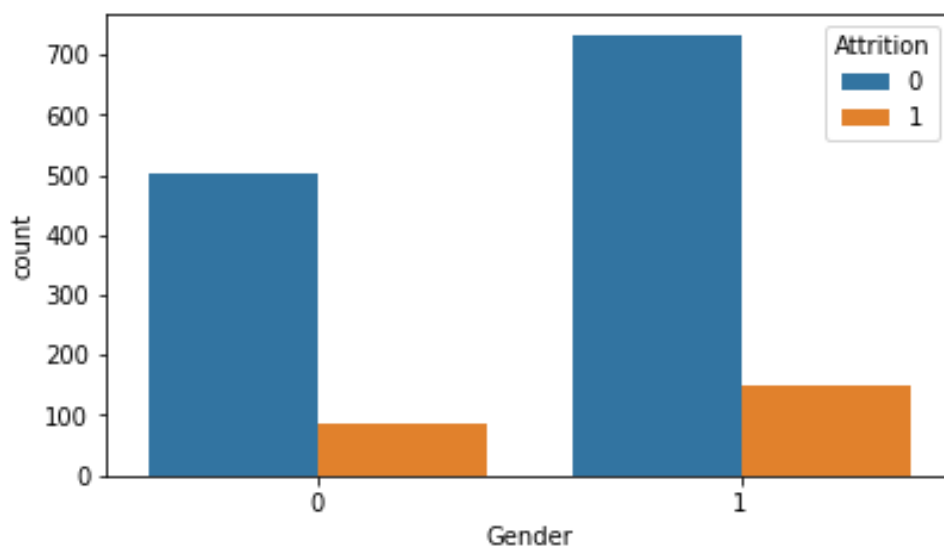
In **18-21 age group**, young employees are more likely to leave the company. Their attrition proportion to their age group is approximately 53.7% (22 out of 41) and that makes up 9% of all attrition (22 out of 237).

If we evaluate overall attrition number in the company, **26-35 age group's attrition number** is the highest comparing to other age groups. In this age group, we have 19.1% of employee attrition (116 out of 606). That makes up approximately 49% of all attrition in the company (116 out of 237).

**35-60 age group** generally prefers to secure their job in the same company.

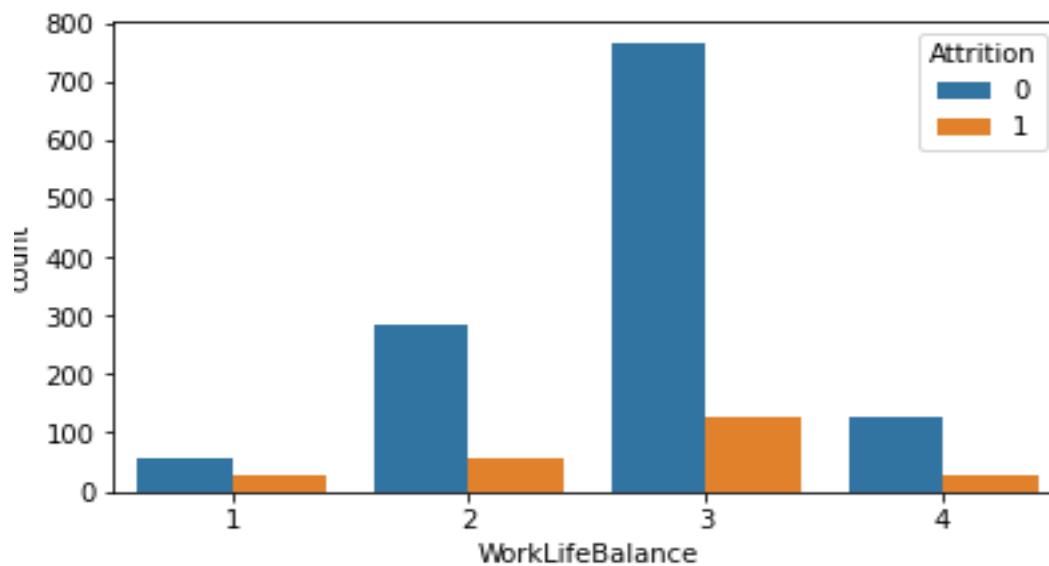


## ii. Gender:



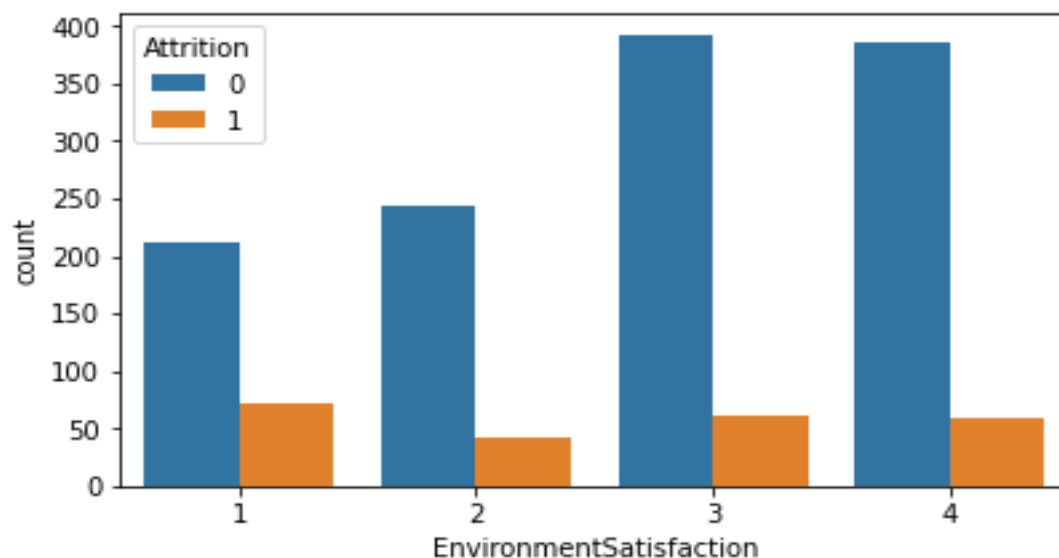
**Male employees** are more likely to leave the company than **female employees**.

iii. **Work Life Balance:**



In general, work life balance is satisfactorily good throughout the company. But we have the highest attrition number and percentage throughout the company. Besides, bad work life balance group has highest attrition percentage in its individual group.

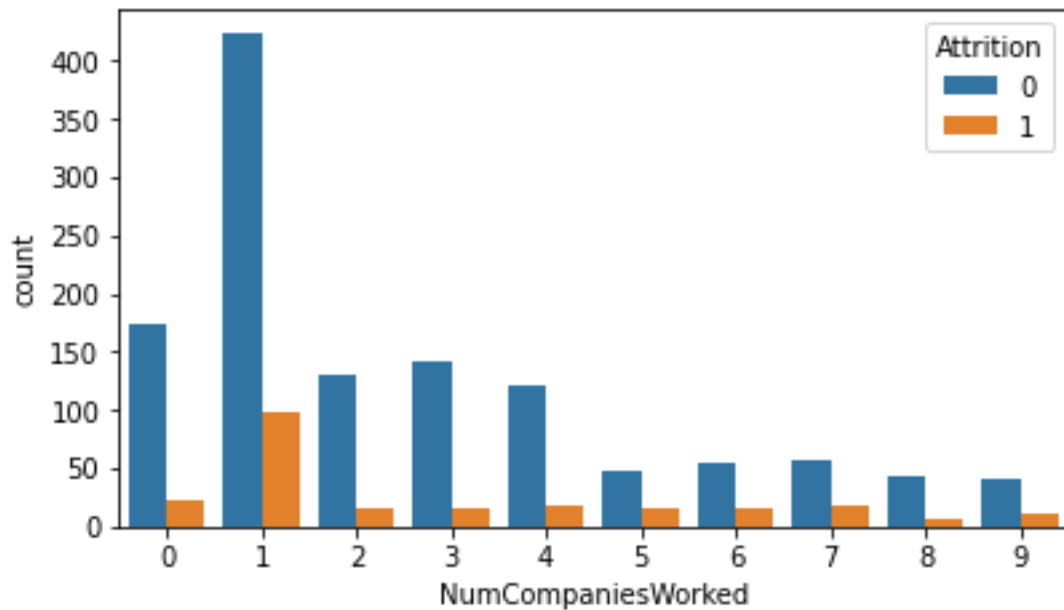
iv. **Environment Satisfaction**



As it may be expected, there is a high attrition rate in the **low satisfaction environment**. That composes the 30.4 % of the whole company's attrition.

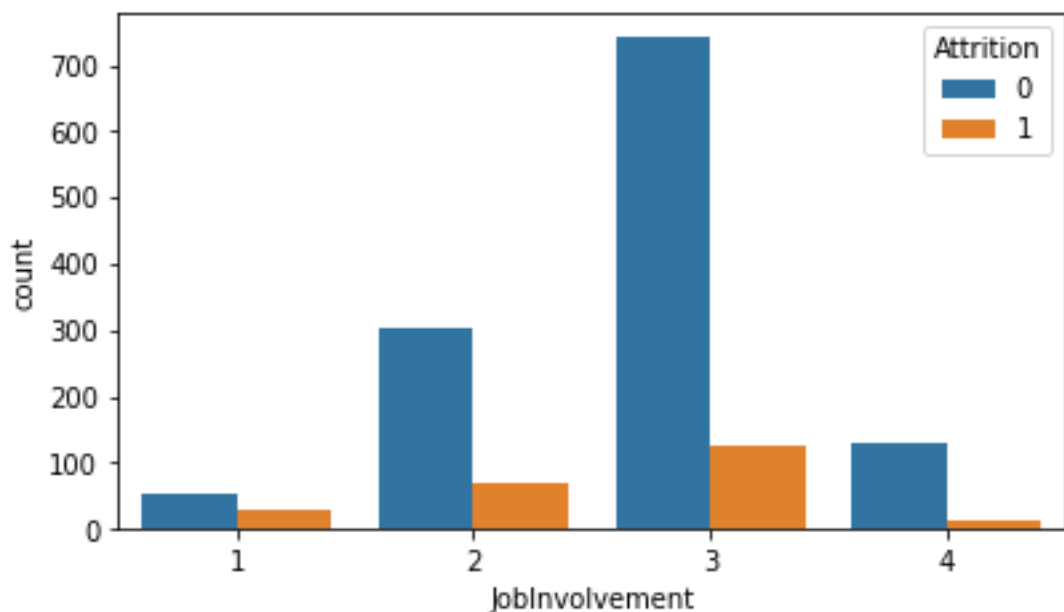
Shockingly, in the **high and very high satisfaction environment**, there are still 13.7 % of the these each group's employees leave the company. That attrition composes of the 51.5 % of the whole company's attrition. This result might tell us that environment satisfaction is not the one of the main reasons for attrition in the company.

v. **Number of Companies Worked:**



35.4% of employees have **one company experience** before current company, and they are more likely to leave the company(18.8%). 21% of employees **worked in 5 and more companies** before this company and have 21.8% attrition.

#### vi. **Job Involvement:**

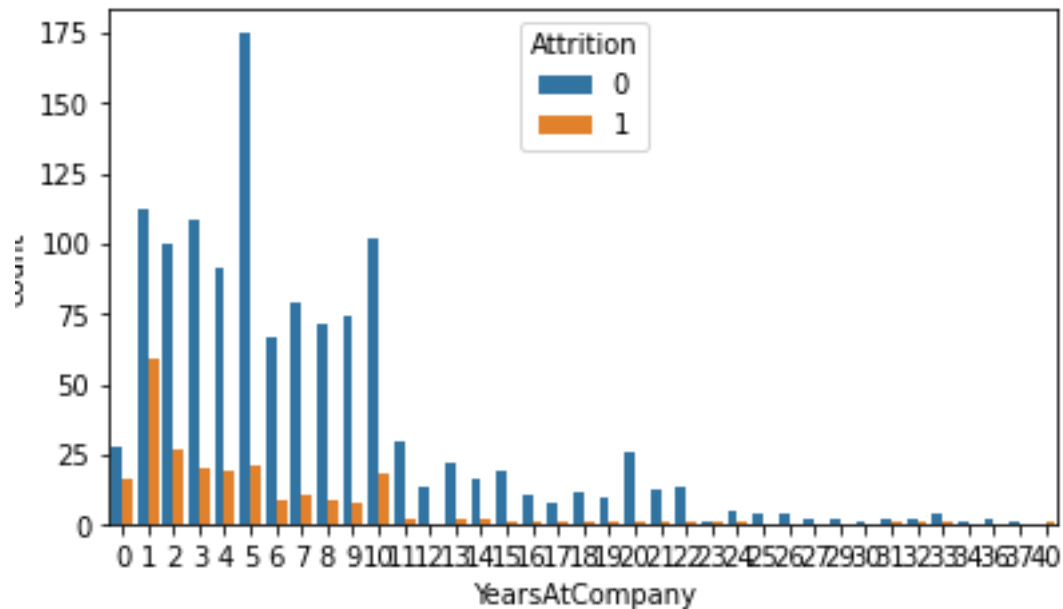


59% of all employee's job involvement in the company is in the **high** category(868 out of 1470). The highest attrition number is also observed in high job involvement category. 125 employees in this group, which composes the 52.7% of all attrition, left the company. But that is only 14.4% of high job involvement category.

**Medium** job involvement category is following the **high** category group in attrition number with 71 employees.

**Low** job involvement category has the highest employee leaving proportion within individual category when it is compared to the other categories. 33.7% of **Low** Job involvement group left the company.

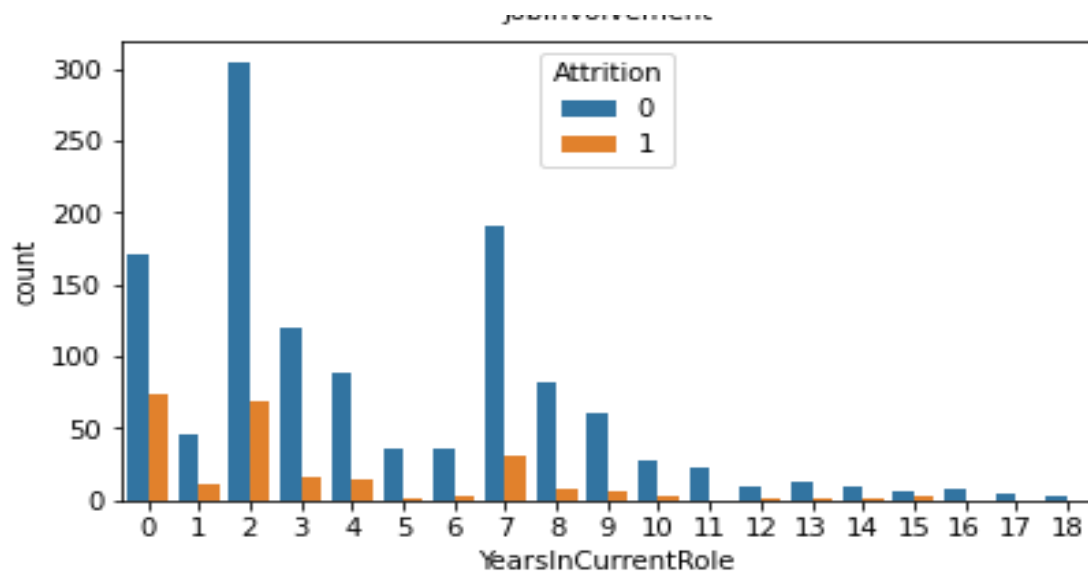
vii. **Total Working Years:**



**Employees who have one year or less working experience** are more likely to leave the company and compose the 18.98% of all attrition throughout the company. In addition to that, **employees who have 6-10 years' experience** have also second highest attrition percentage throughout the company and it compose the 38.4% of all attrition.

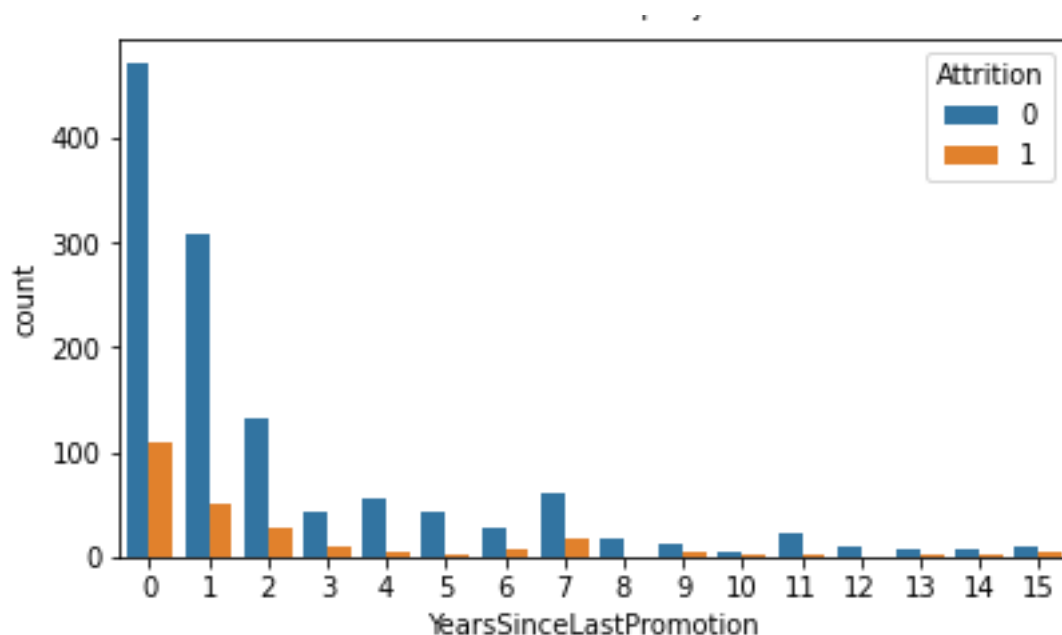


viii. **Years in Current Role:**



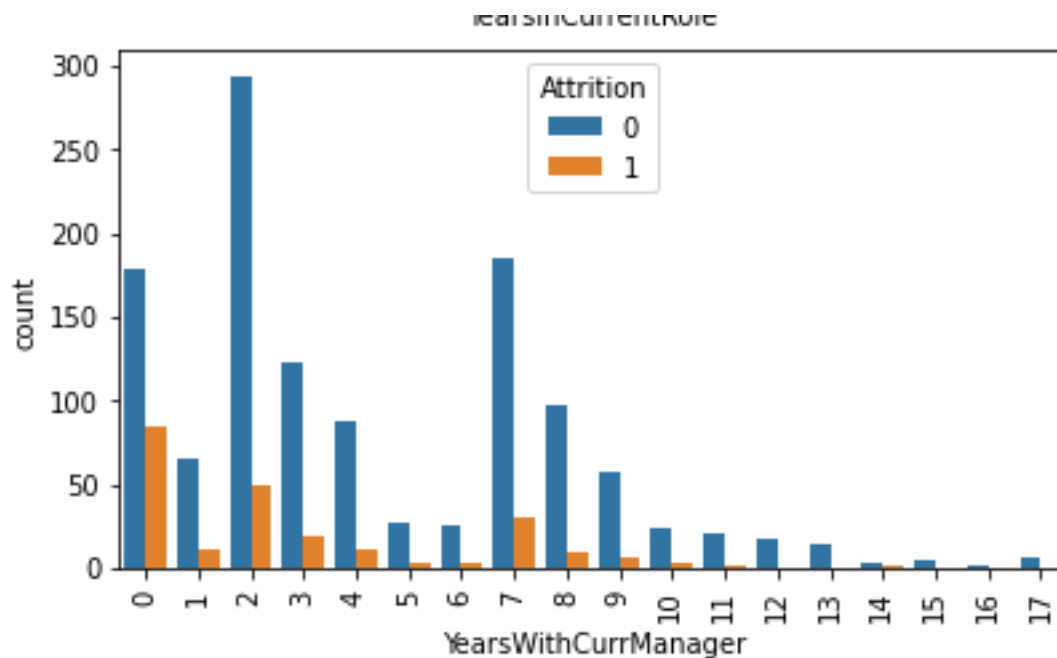
Employees who don't fulfill their first year and in their first year in their current role are more likely to leave the company. That might be result of challenge or not satisfied with the current role. Employees who have 2-5 years' experience in that company compose of the maximum attrition percentage and number in the company. Besides that, after years in current role, employees are willing to leave the company. That might be result of looking for better opportunities in other companies.

ix. **Years Since last Promotion:**



Employees who don't fulfill his one year since the last promotion in the company are more likely to leave the company(46.4% of all attrition). And employees who have one- and two-years' experience in the current company since the last promotion have the highest attrition number after the above group in the company respectively. 7 years since last promotion has the highest attrition in its individual group.

x. **Years With Current Manager:**

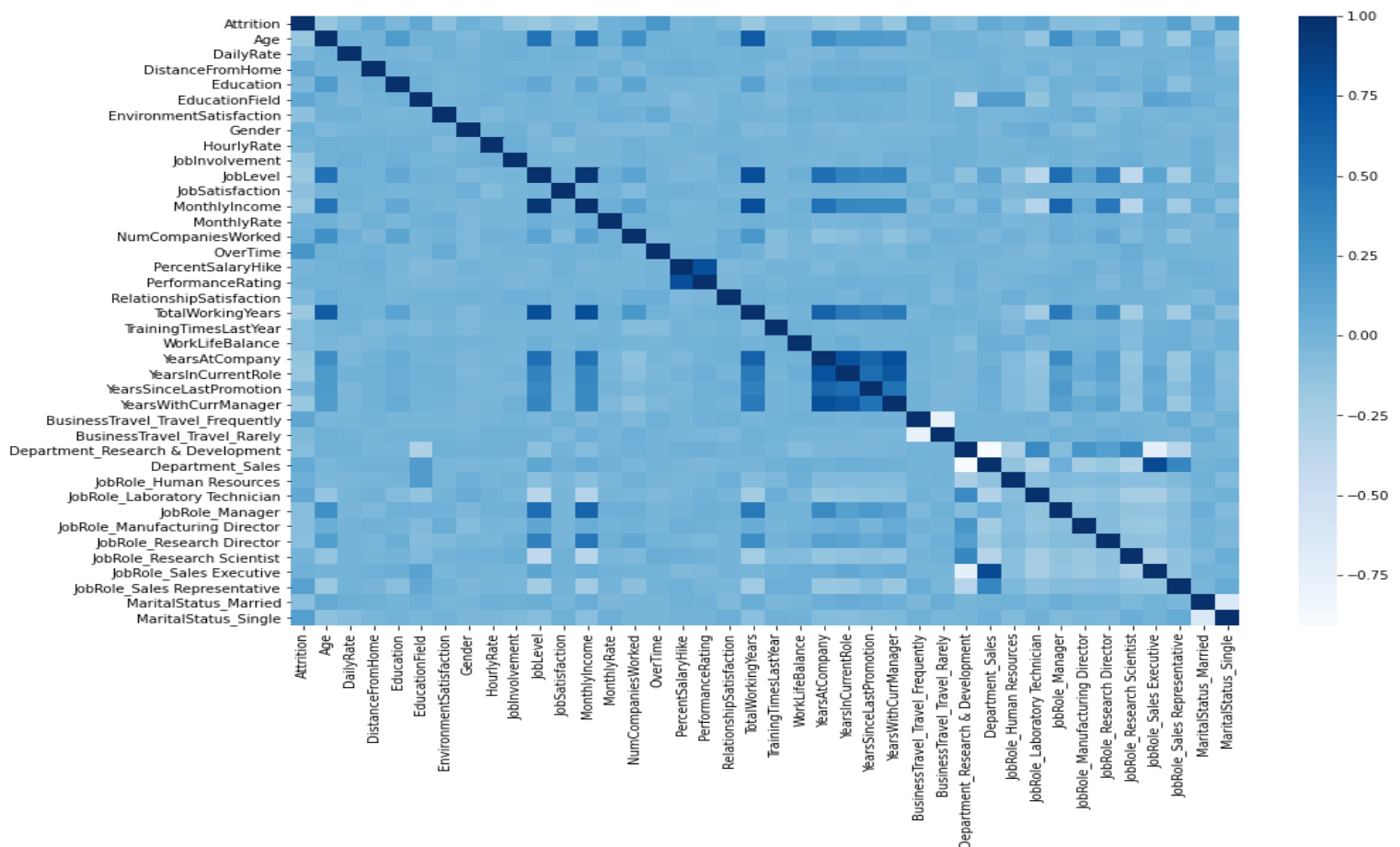


Most of the employee quit the company before completing their first year with their current manager. Other group who leaves the company most is the ones who work two years with current manager.

**Feature/Variable Relationships:**

In this section, I looked at how variables related to each other. There are various methods/visualizations for this. I used correlation matrix (heat map) for this purpose.

Correlation means association - more precisely it is a measure of the extent to which two variables are related. There are three possible results of a correlational study: a positive correlation, a negative correlation, and no correlation. A positive correlation is a relationship between two variables in which both variables either increase or decrease at the same time. An example would be height and weight. Taller people tend to be heavier. A negative correlation is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. An example would be height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature). A zero correlation exists when there is no relationship between two variables. For example, there is no relationship between the amount of tea drunk and level of intelligence.



Based on the fact which is given strength of correlation chart, we can identify the features which have strong, moderate, weak and zero correlations between each other. I will just outline the strong and moderate correlations here.

### Features which have strong correlations:

Percent Salary Hike and Performance Rating,

Total Working Years, Monthly Income and Job Level,

Years at Company, Years with Current Manager, and Years in Current Role,

## Summary

There are 1470 employees in the company and 16% of them left the company. We have some data about employees to examine the attrition reasons. To sum our exploratory data analysis;

**Young employees (18-25 years old)** compose the 8.4% of the company and they are more likely to leave the company than other age groups. **As the employees get older, their attrition percentages drop.**

61.2% of employees has **environment satisfaction at very high/high level** in the company. 19.3% of employees has **low environment satisfaction** and 25.3% attrition.

**Male employees(60% of company)** are more likely to leave the company than **female employees**.

68.8% of employees has **high/very high job involvement** in the company. As the **job involvement** increases, the attrition percentage decreases respectively.

35.4% of employees have **one company experience** before current company, and they are more likely to leave the company(18.8%). 21% of employees **worked in 5 and more companies** before this company and have 21.8% attrition.

The higher percent salary hike is, the more employees are likely and willingly to stay in the current company.

84.2% of employees have **good or better work life balance**, and 16.7% and 14.2% attrition respectively. 5.4% of employees have **bad work life balance** and the highest attrition percentage(31.3%).

As the **years at company and current role** increase, attrition percentage decreases in the company.

39.5% of employees leaves the company **before fulfilling one year since their last promotion**. 24.3% of employees **fulfilling one year since their last promotion** has the 13.7 attrition percentage. **7 years since last promotion employee group** (5.2%) has the highest attrition(21.1%) in its individual group.

Most of the employee(17.9%) quit the company **before completing their first year with their current manager**(32.3% attrition). Other group who leaves the company most is the ones who work **two years with current manager**(23.4% of employees with 14.5% attrition).

## Model Building And Fitting the Data

Significant Variables identified in linear models are also used in Random forest

Below table provides a snapshot of the various models which the business can choose from based on the pros and cons of each model.

Sr. No	Model Name	Accuracy	Benefits	Trade-offs
1	Logistic_Regression_Model	83.06	High interpretability Significant variables can be easily identified	May give lower accuracy
2	Decesion_Tree_Model	75.51	Significant variables are automatically selected.	May give lower accuracy without pruning.
3	Random_Forest_Model	84.89	Good Accuracy	Minimal / lower interpretability
4	ADA boost	85.5	Improves accuracy by combining weak learners	Space and time taken is more
5	XGB	84.49	Highest Accuracy	Minimal / lower interpretability

## **Suggested Actions:**

It's not sensible to focus on every employee who wants to leave because it costs time and energy for human resources management department. HR department need to focus on:

- Improving the work conditions

Provide an option for the employee's to work from home, on a flexible schedule, or in an office with an ergonomic workspace, they will be more satisfied with their work and more likely to achieve a healthy work-life balance.

- Offer modest salaries and perks

To maintain the critical employee's company need's to offer equitable and modest salaries. You can also give added perks like flexible schedules, travel discounts etc.

- Employee Engagement

When you have talented employee's we need to find ways that you can help expand the employee's skill set, so that their involvement in the job increases. If their involvement is low, they will get bored and think that they are not growing within the organization