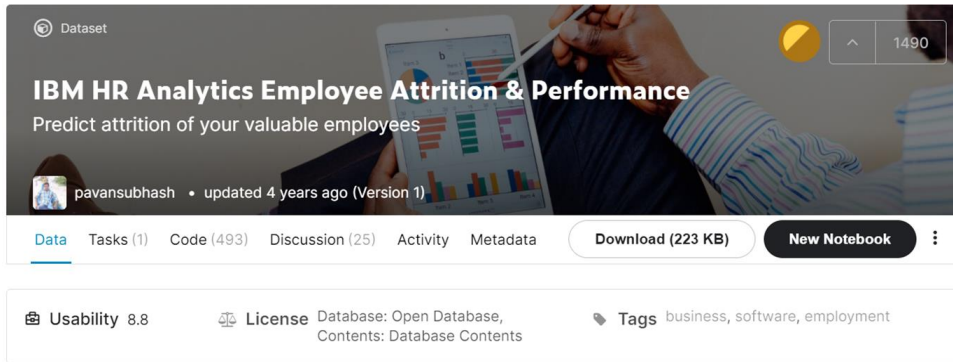


# **IBM HR Analytics Employee Attrition & Performance**

**Data Science Project**  
**by: Angelin Grace Wijaya, Agarwala Pratham, Krishna Shivangi**

# SOURCE

- Kaggle Dataset by **pavansubhash**
- **Fictional** dataset created by IBM data scientists
- Contains 35 columns and 1470 records
- No missing values



The screenshot shows the Kaggle dataset page for 'IBM HR Analytics Employee Attrition & Performance'. The page features a header with the dataset title and a subtitle 'Predict attrition of your valuable employees'. Below the header, there is a section for the creator 'pavansubhash' with a profile picture and a note 'updated 4 years ago (Version 1)'. The main content area includes tabs for 'Data', 'Tasks (1)', 'Code (493)', 'Discussion (25)', 'Activity', and 'Metadata'. A 'Download (223 KB)' button and a 'New Notebook' button are also visible. At the bottom, there is a section for 'Usability 8.8', 'License Database: Open Database, Contents: Database Contents', and 'Tags business, software, employment'.

<https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset>



# Business Objective of our Project

- **HR Analysis tool** for companies especially those with large workforce.
- Help companies to be prepared for future employee-loss
- To find possible reasons for employee attrition, in order to prevent valuable employees from leaving.
- Predicting employee's Performance Rating and hence distributing the employees into two classes  
( 0 : Low Performance Rating, 1 : High Performance Rating)
- Accordingly, predict the employees "Percent Salary Hike" for the next working term.



# DATA CLEANING

Two of the columns in the employee dataset are labelled:

“**Standard Hours**” and “**Employee Count**”.

In “Standard Hours”, we found that all the records have a value of 80. What does this mean?

This standard hour is also called a **9/80 work schedule**. Here, employees work for 80 hours over a period of 9 days.

In “Employee Count”, we found that all the records have a value of 1. What does this mean?

Employee count is also known as **headcount**. It is the number of employees working in the company, where each employee, whether full-time or part-time, is counted as 1.

As these columns all have the same values, we decided to drop them while keeping in mind of what each of them represents.

# DATA CLEANING

Two of the columns in the employee dataset are labelled:

“**Over18**” and “**NumCompaniesWorked**”.

In “Over18”, we found that all the records have a value of “Yes”.

We found that there are actually 8 employees who are 18 from the “Age” column. This means that the “Over18” column has mismatched values, or that this column actually means employees who are 18 and above.

We kept this column and **change the records for those who are aged 18 to “No”** in their “Over 18” column.

In “NumCompaniesWorked”, we found that there are records with **zero companies worked**.

We found that for these employees, they all have 1 year where they worked not for a company (ex. part-time).

As 18 year-old employees, who have zero years of work experience, have one as number of companies worked (their current one), we also decided to **change the records of number of companies worked as zero to one**.

# Exploratory Data Analysis

---

# JOBS

## Sales department:

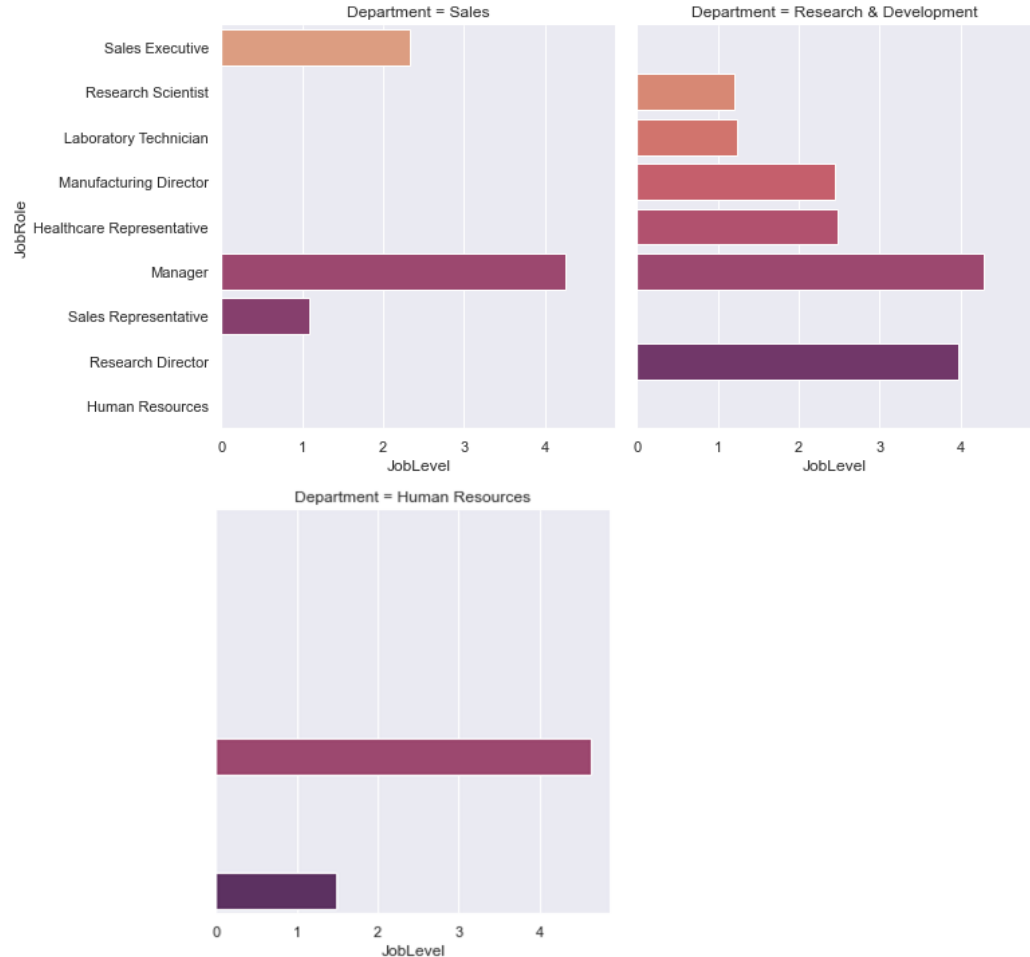
- Sales Executive
- Sales Representative
- Sales Manager

## Research & Development department:

- Research Scientist
- Laboratory Technician
- Manufacturing Director
- Healthcare Representative
- Research Director
- Research Manager

## Human Resources department:

- Human Resources
- Human Resources Manager

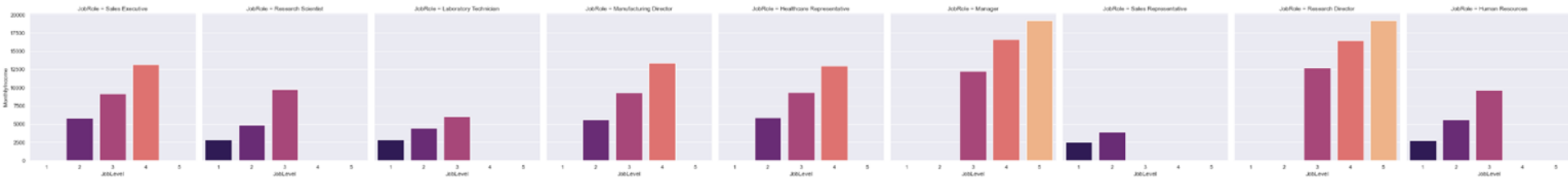
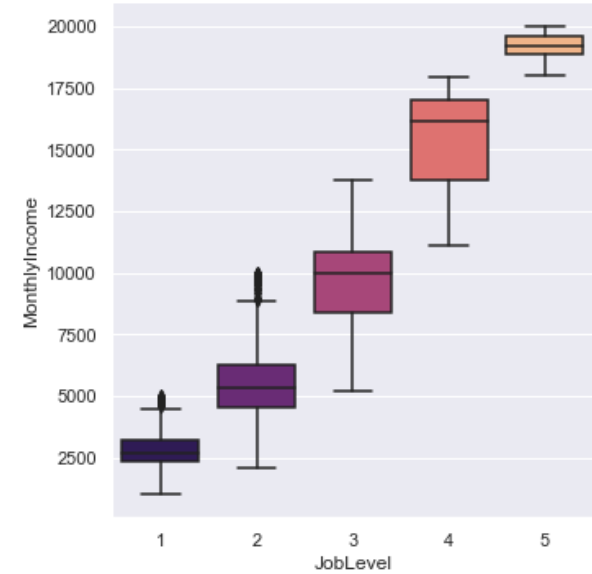


# JOB LEVELS

It is defined as the classification of authority and/or salary levels of positions in an organization.

Ex. senior executives, executive, middle management, advisors, associate

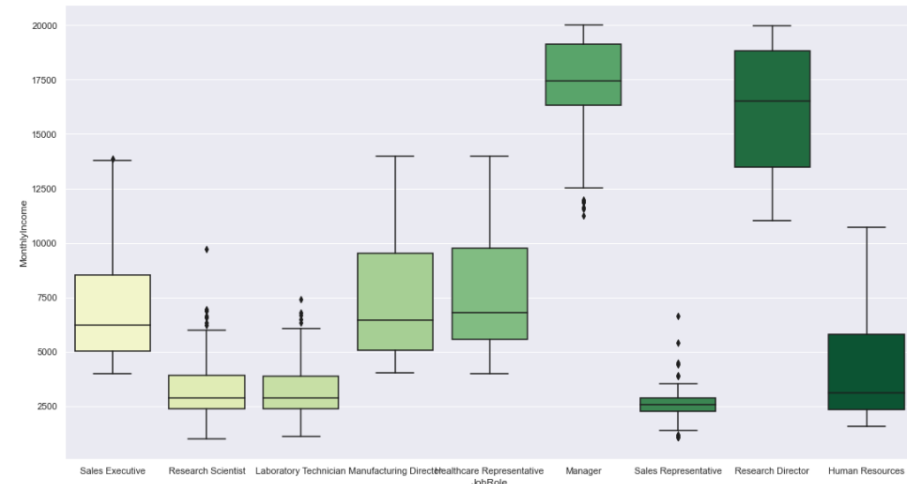
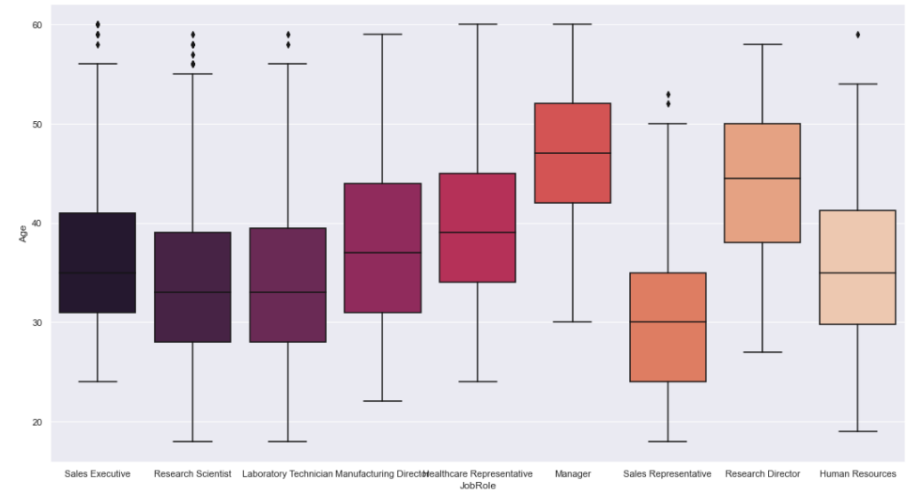
Different job levels have different monthly income.



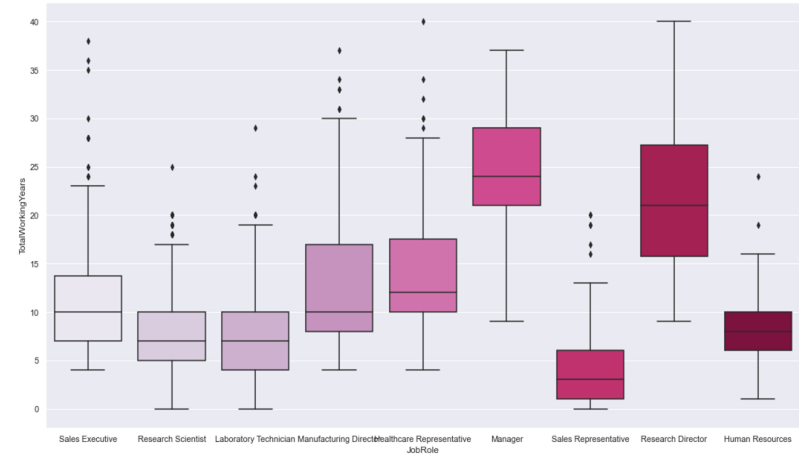
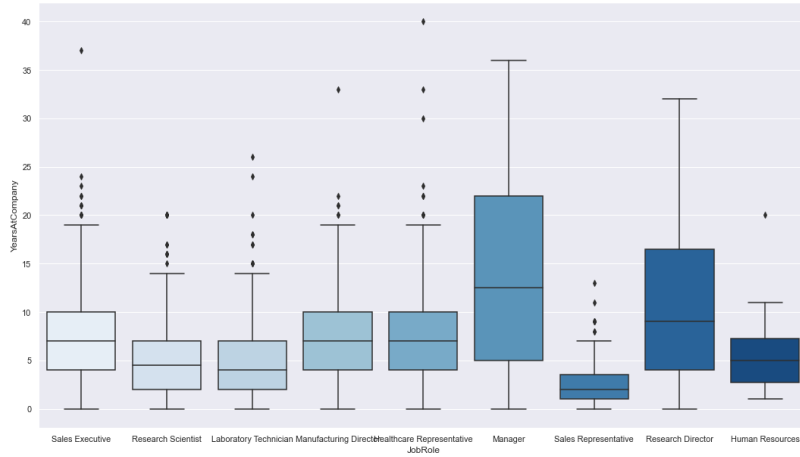


# JOB ROLE & AGE

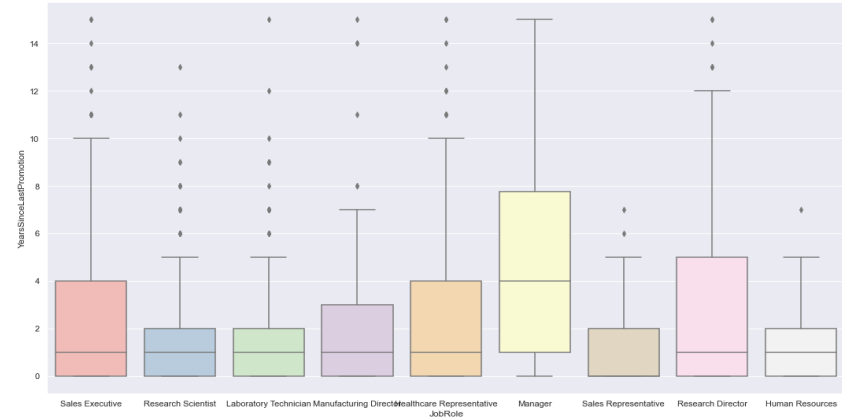
- Managers and Research Directors have the oldest median age.
- Managers and Research Directors are also earning the highest median monthly income
- Sales representatives are the youngest and they are receiving the lowest monthly income



# JOB ROLE WITH TOTAL WORKING YEARS & YEARS IN CURRENT ROLE



- Managers and research directors have the highest total working years and years at current company
- Sales representatives have the lowest total working years and years at current company
- Years since last promotion is more or less the same for all the job roles other than managers



# Predicting Attrition of Employees

---

# WHAT IS ATTRITION?

Attrition is the **decrease** in the number of employees over time as they **leave** due to personal reasons or **retire**.

**Low attrition rate** is desired by many companies because it implies that employees are **satisfied** with their workplace and the company does not have to train new recruits.

## Disadvantages of Attrition:

1. Decrease Overall Performance
2. Difficulty in Managing Task *(that was left by the employee)*
3. Increased Cost *(must hire and train new employee)*
4. Lack of Knowledgeable/Experienced Employee
5. Losing Out on Employee Development Plan
6. Negative Image



# DECISION TREE

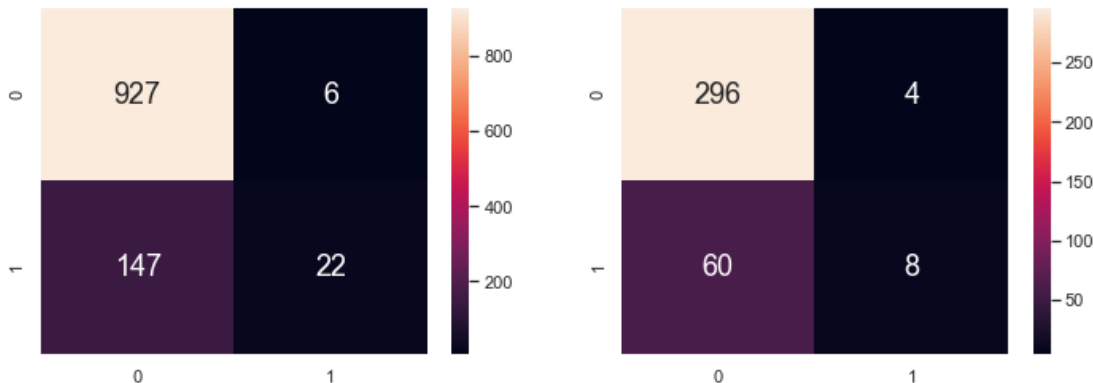
We found that our dataset is **imbalanced**.

The ratio of the number of people with attrition to without attrition is 1233 : 237

The confusion matrix above is before balancing.

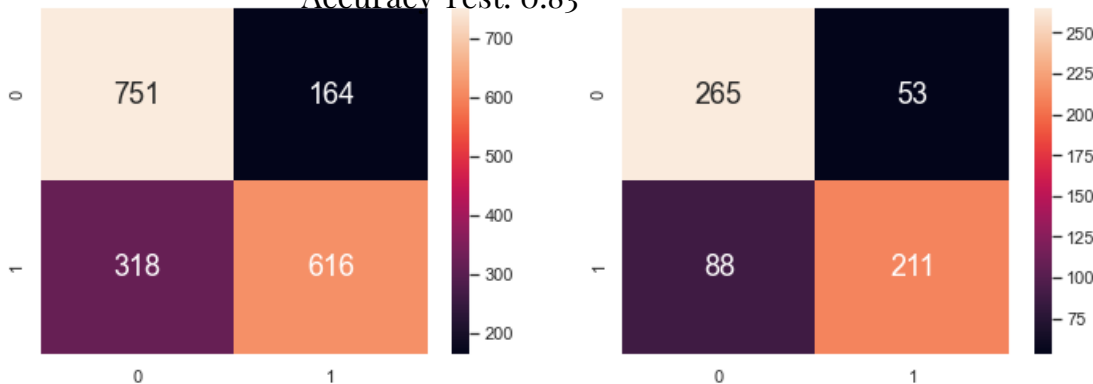
Those who actually have attrition are being classified as having no attrition.

Below is the matrix after balancing.



Accuracy Train: 0.86

Accuracy Test: 0.83



Accuracy Train: 0.74

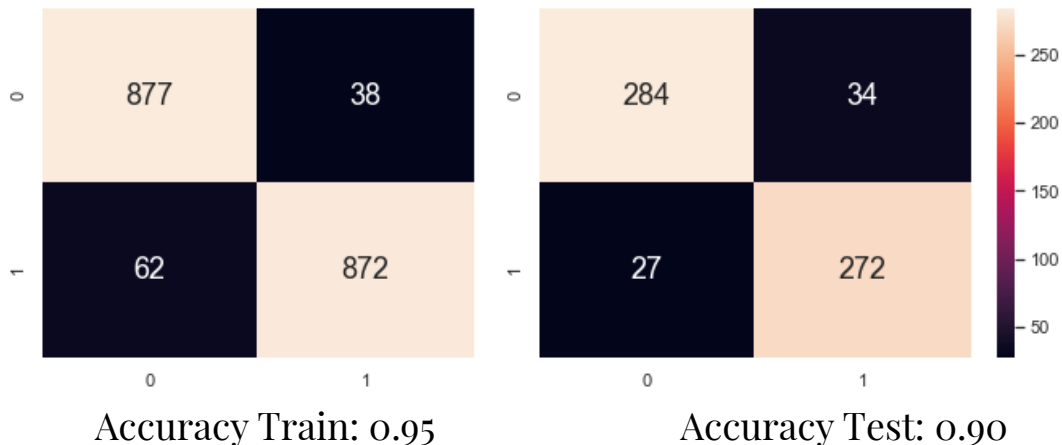
Accuracy Test: 0.77

# RANDOM FOREST

We used **GridSearchCV** to find the ideal number of trees and depth for our random forest.

Result: `{'max_depth': 7, 'n_estimators': 100}`

Below is our confusion matrix for random forest:



Cross-validation is performed by the GridsearchCV.

We pass the whole dataset to it without splitting it and by default it performs 5-fold cross validation

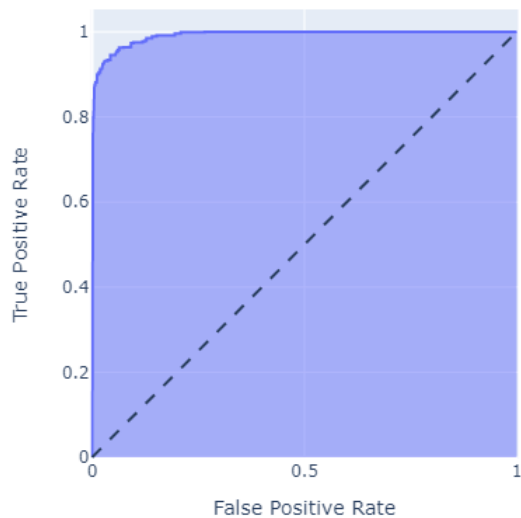
## Why random forest?

A single decision tree like the one in the previous slide brings a much lower accuracy than forest.

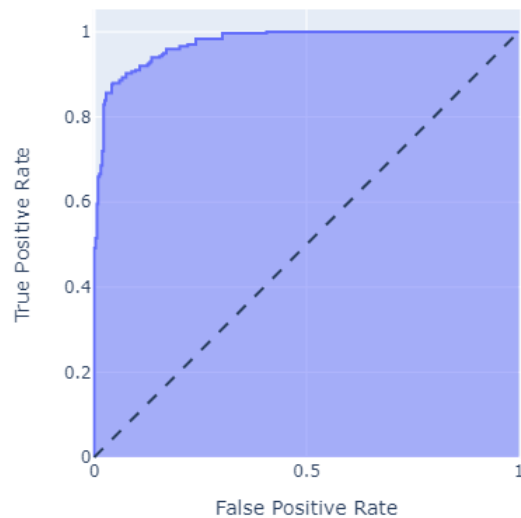
In random forest, we can plot multiple trees that learn from one another's mistake (or wrong prediction) to give a better accuracy.

# ROC CURVES

ROC Curve for Train (AUC=0.9919)



ROC Curve for Test (AUC=0.9737)



We also found the Best Threshold to be 0.498017 and optimal G-Mean as 0.951

# VARIABLE IMPORTANCE

1. Overtime

2. Monthly Income

3. Years At Company

4. Job Level

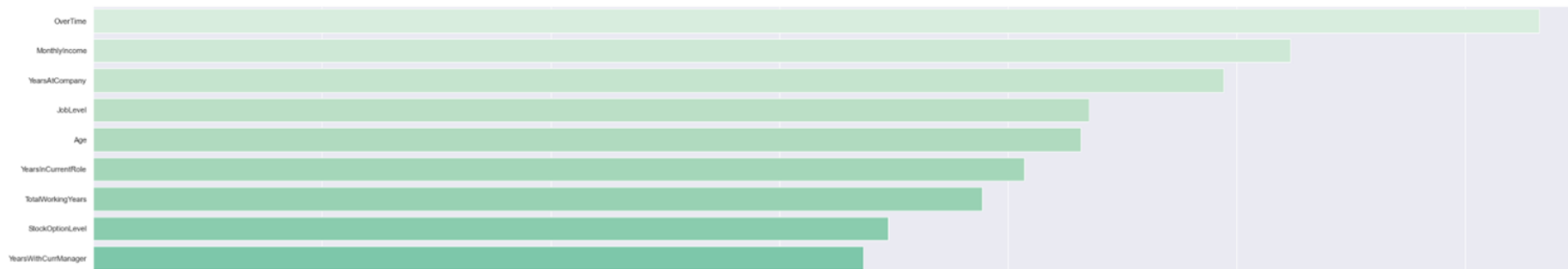
5. Age

6. Years In Current Role

7. Total Working Years

8. Stock Option Level

9. Years With Current Manager



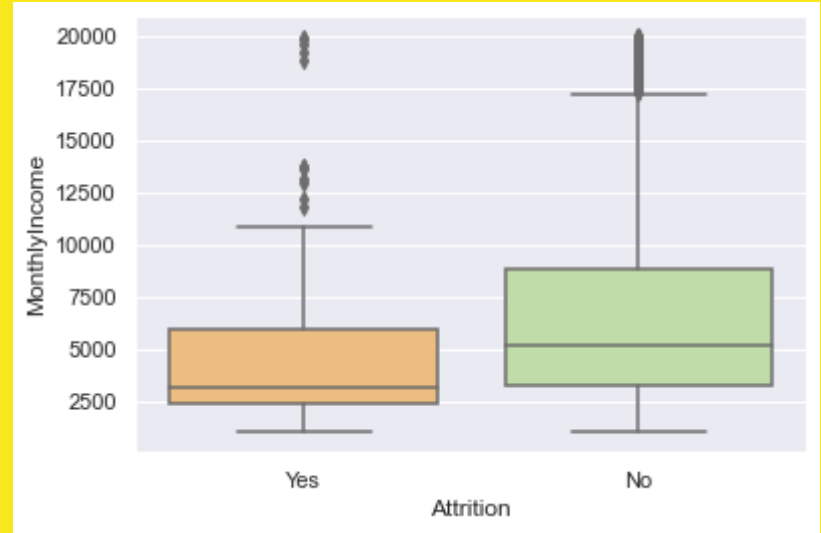
*Train Test Split has been fixed with random state of 1.*

*Without random state, these 9 variables are still in the top 9 importance but may be in different ranks.  
(ex. Monthly Income may become more important than Overtime, but is still within the 9 ranks)*



# Attrition and Monthly Income

Employees with attrition have lower monthly income

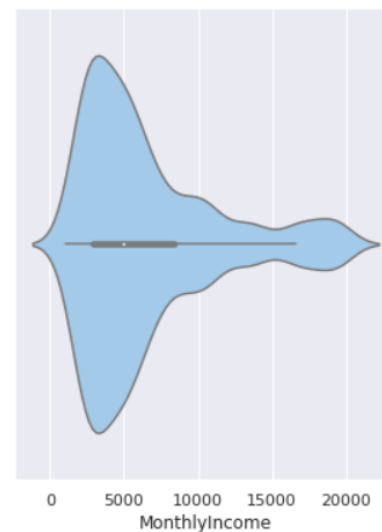
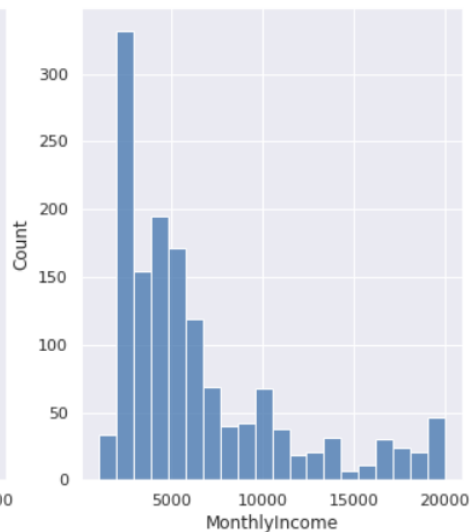
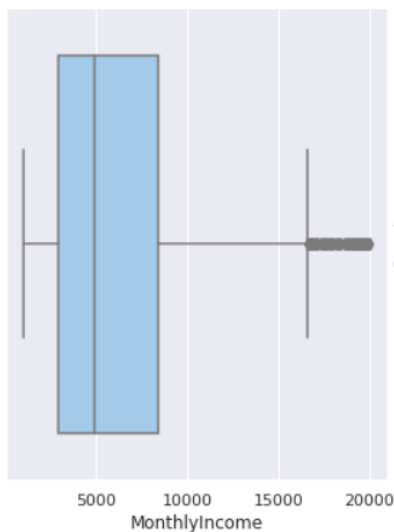


Attrition also has a  $-0.159840$  correlation with Monthly Income

# MONTHLY INCOME- MORE INSIGHTS

First of all, Monthly Income is a continuous data, so we performed extensive EDA to study the patterns in Monthly Income .

1. **RFECV** : Recursive Feature Elimination from `sklearn.feature_selection` and for CV we will use KFold validation.
2. **Greedy Search** : Iteratively removing variables from the dataset until there are only five variables left excluding the response variable(MonthlyIncome) and criterion is the accuracy score of the model . It removes the variables that least increase the accuracy score .



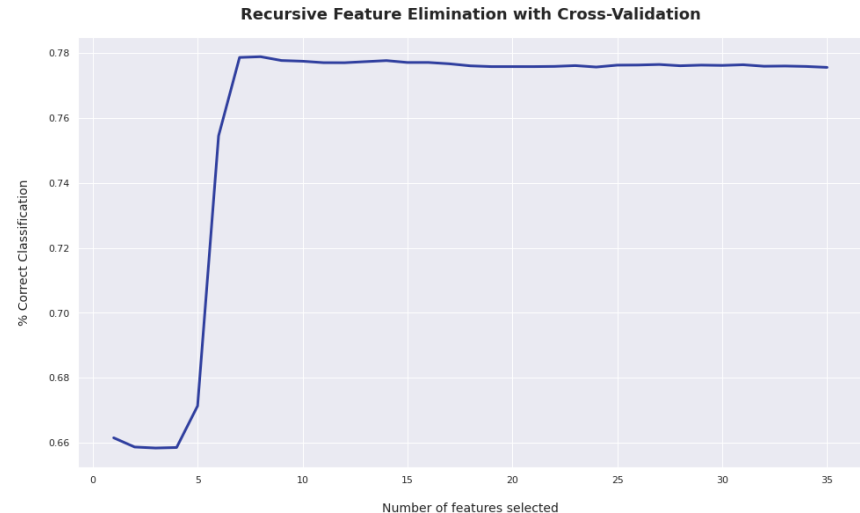
# MONTHLY INCOME – RFECV

Using RFECV(Recursive Feature Elimination CV\ross Validation) from sklearn.feature\_selection we performed the following the steps:

1. We label-encoded the variables which have some kind of order or have classes 2 or less, and for rest we did One-Hot Encoding.
2. Removing correlated features is done because we don't want highly correlated features in our dataset as they provide the same information — one is enough. For this step we wrote a simple algorithm. We iterated through the entire correlation matrix and removed variables with correlation with other variables exceeding 0.75.
3. We declared two variables — X and target where first represents all the features, and the second represents the target variable. Then we'll use an Machine learning Model (We are using **LinearRegression**). Then we ran **RFECV**. Here are the arguments:
  - a. Estimator : LinearRegression()
  - b. Cv : KFold(10)
  - c. Scoring : 'Accuracy'
4. Now seeing the graph the optimal of feature is **8** to get Best accuracy for LinearRegression Model.

Variables Short-listed after using RFECV:

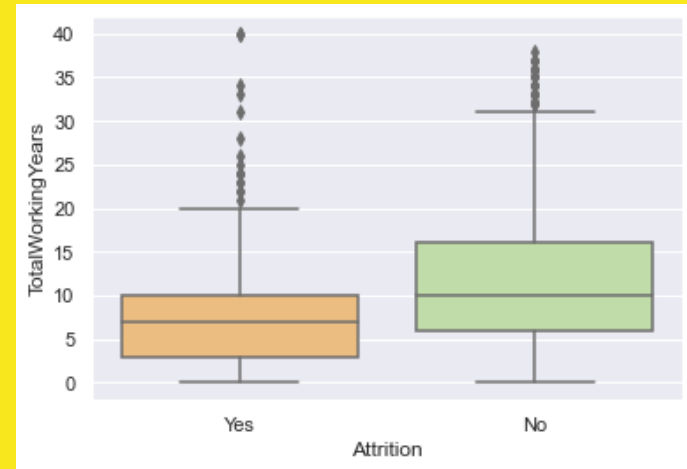
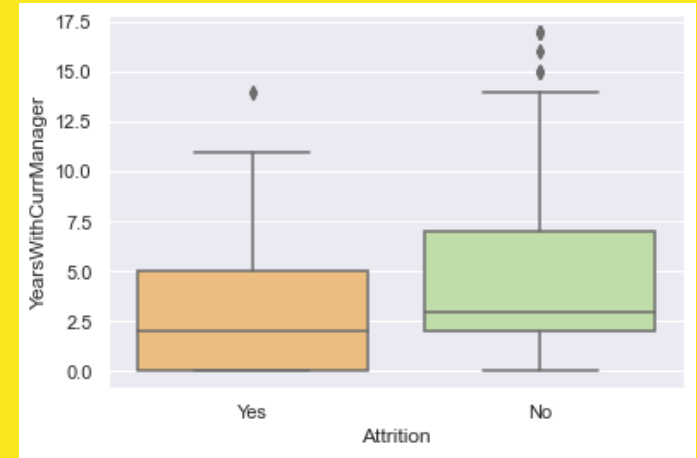
- |                            |                                    |
|----------------------------|------------------------------------|
| → <b>TotalWorkingYears</b> | → <b>BusinessTravel_Non-Travel</b> |
| → <b>Attrition</b>         | → <b>EducationField_Human</b>      |
| → <b>Department</b>        | → <b>Resources</b>                 |
| → <b>JobRole</b>           | → <b>EducationField_Marketing</b>  |
| → <b>Over18</b>            |                                    |

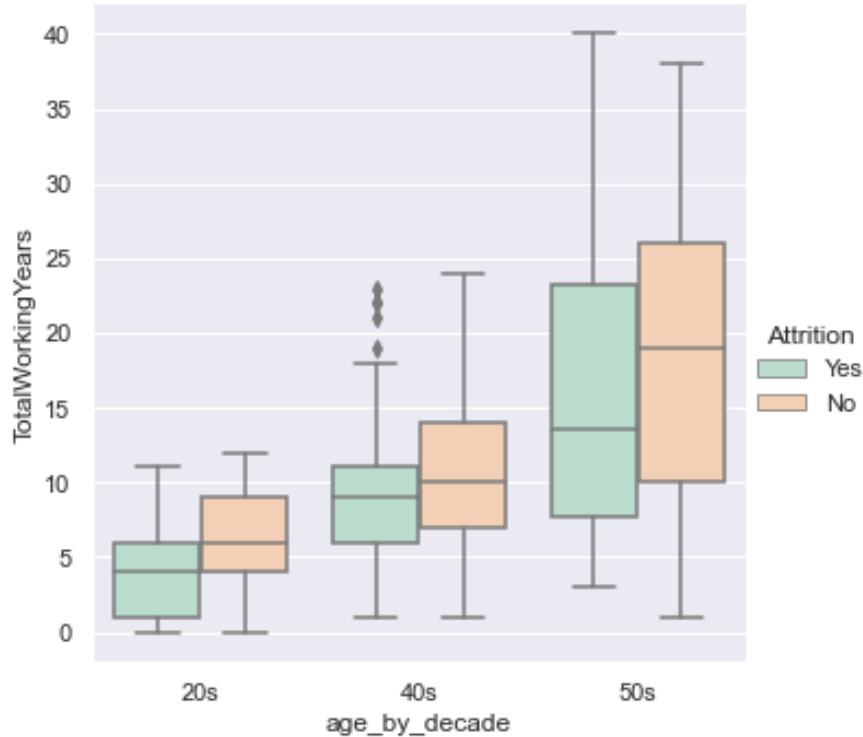


# Attrition and Years (1)

Employees with attrition have lower years with current manager and total working years

Attrition has a  $-0.156199$  correlation with Years With Current Manager and a  $-0.171063$  correlation with Total Working Years





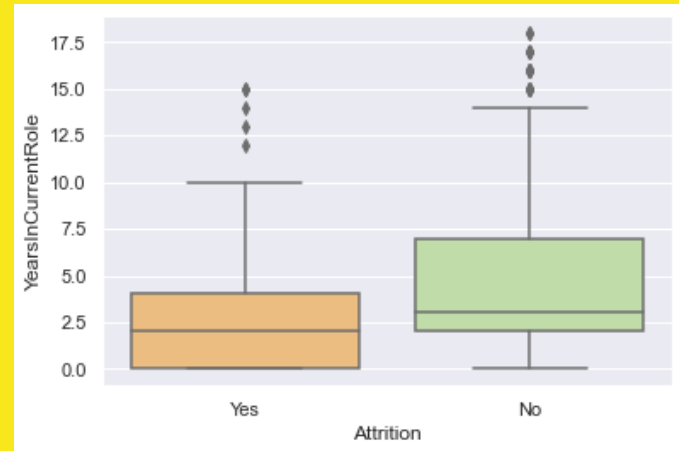
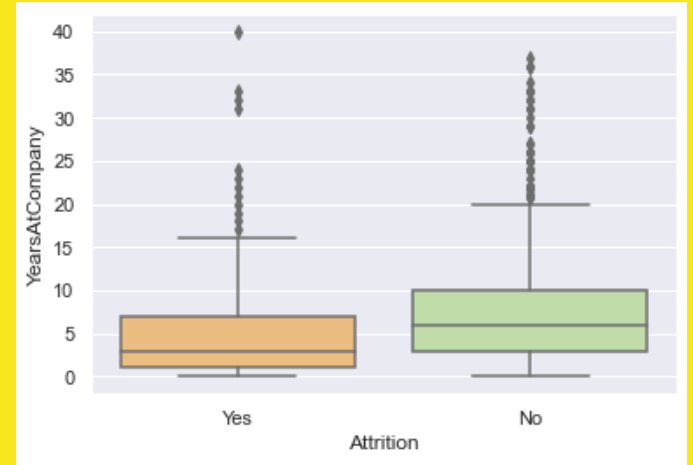
### Age and Total working years:

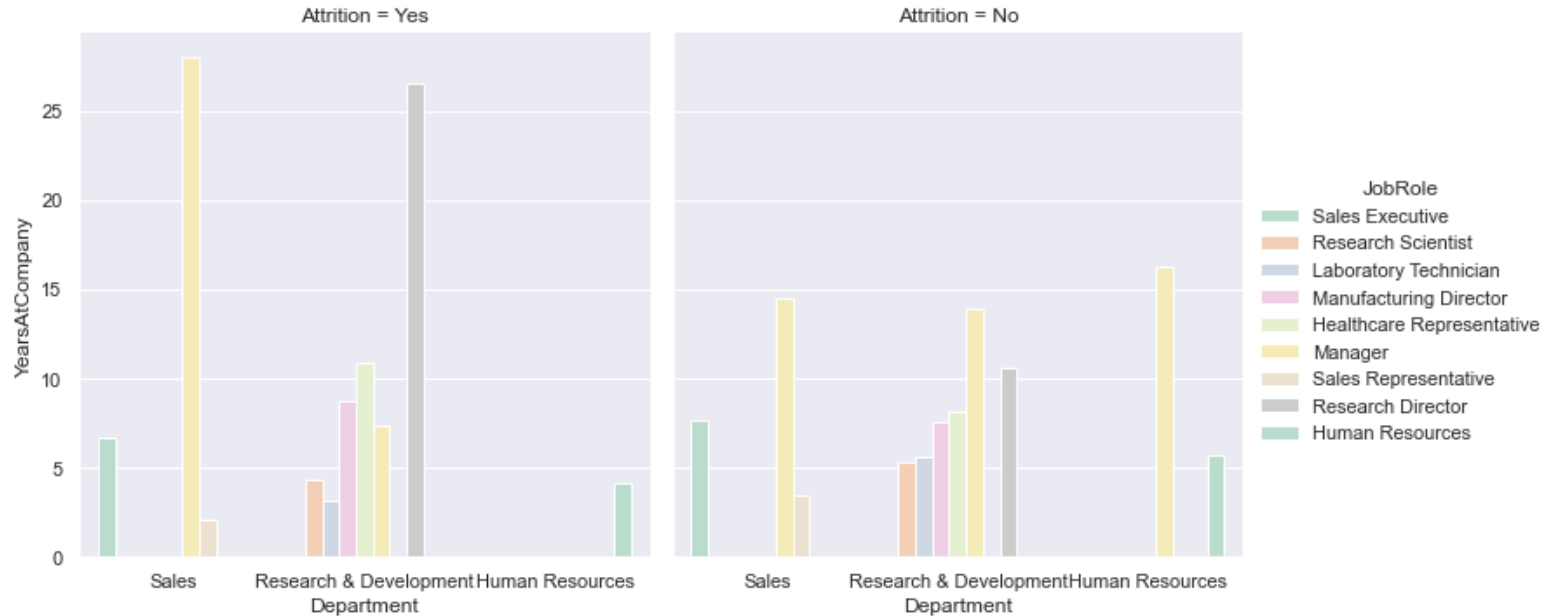
- The age has been classified as this:
  - 20s -> younger employees
  - 40s -> middle aged employees
  - 50s -> older employees
- The correlation between total working years and age is 0.68
- For all three categories, the number of employees who have worked for a shorter period of time are the ones who want to attrite
- As the age increases, the total number of working years increases

# Attrition and Years (2)

Employees with attrition have lower years at company and lower years in current role

Attrition has a  $-0.134392$  correlation with Years At Company and a  $-0.160545$  correlation with Years In Current Role



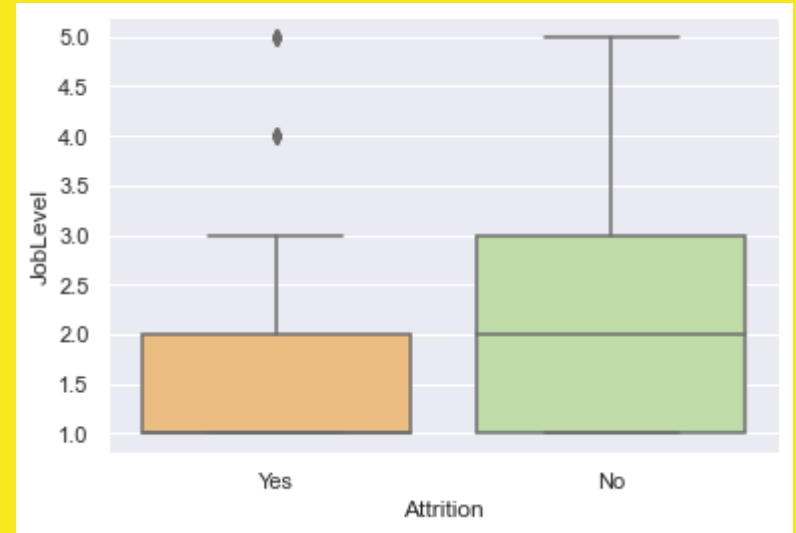


Among attiring employees,

- Sales Manager and Research Director are attiring after longer years at company.
- We know that Managers and Research Directors have the oldest median age and the highest median monthly income. Therefore, it is highly possible that they are attiring due to **retirement**.
- Employees that attire with the least number of working years are sales representatives.
- Other job roles are attiring after shorter working years, and we predict that it is to look for **better job opportunities**.

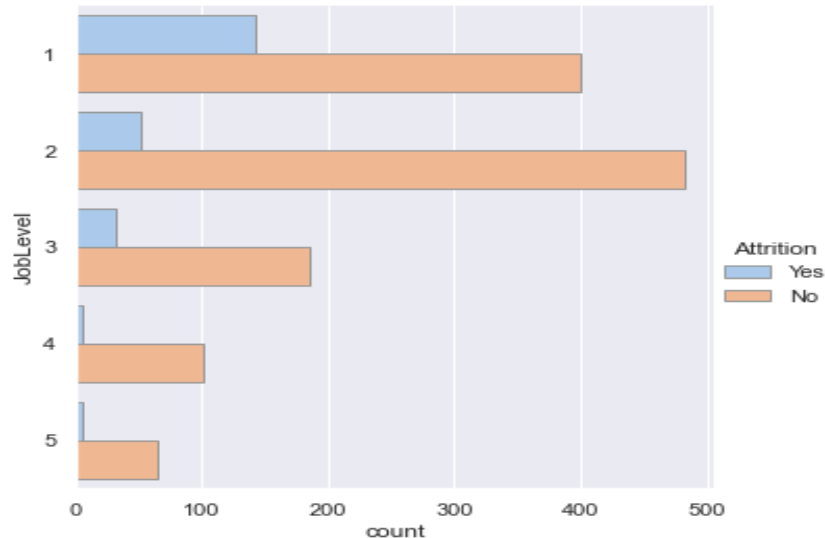
# Attrition and Job Level

Employees with a lower job level are more likely to attrite



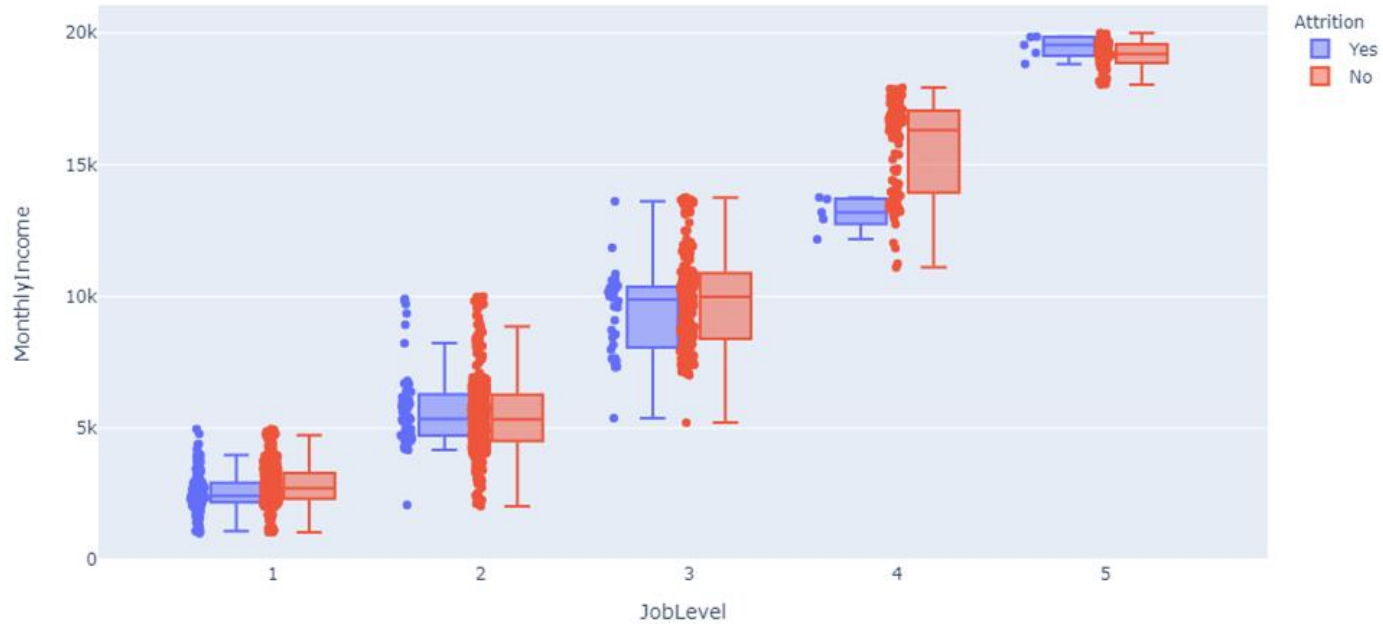
Attrition has a  $-0.169105$  correlation with Job Level





## Job level and Monthly Income

- The correlation between job level and monthly income is 0.95
- Employees with a job level of 5 are either managers and research directors (highest paid)
- Employees with a job level of 1 are sales representatives (lowest paid)
- As the job level increases, the number of employees that want to attire decrease
- Job level of 1 has the highest number of employees that want to attire



From this boxplot, we can see how every job level almost has a corresponding range of monthly income.

The difference in median monthly income (No Attrition - Attrition) is greatest for Job Level= 4 followed by Job Level= 1.

A possible reason for **high attrition rate at Job Level 1** may be that they feel it is **unfair** to be receiving a **lower monthly income** than others in the **same job level**.

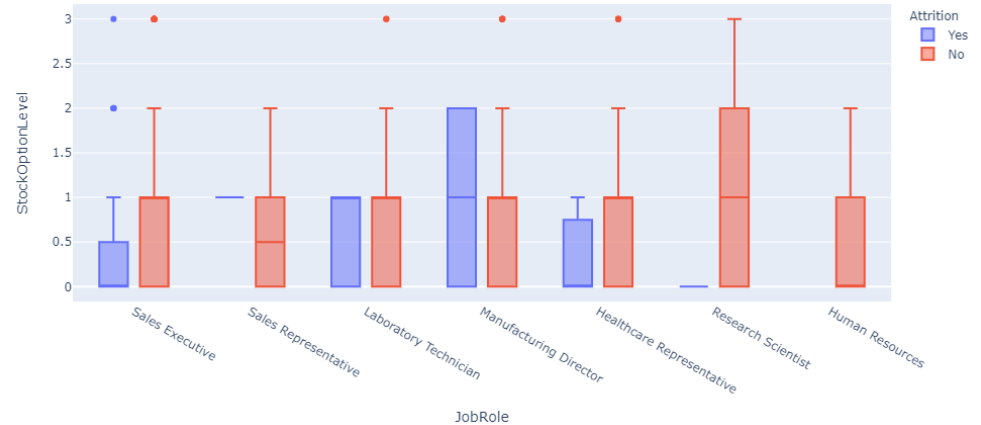
How about Job level 2 then? The monthly income for both are almost equal, so what could be the reason for them attriting?

Also, the difference in monthly income in level 4 is very high, why is the attrition rate lower?

## Possible reason for **high attrition in Job Level 2:**

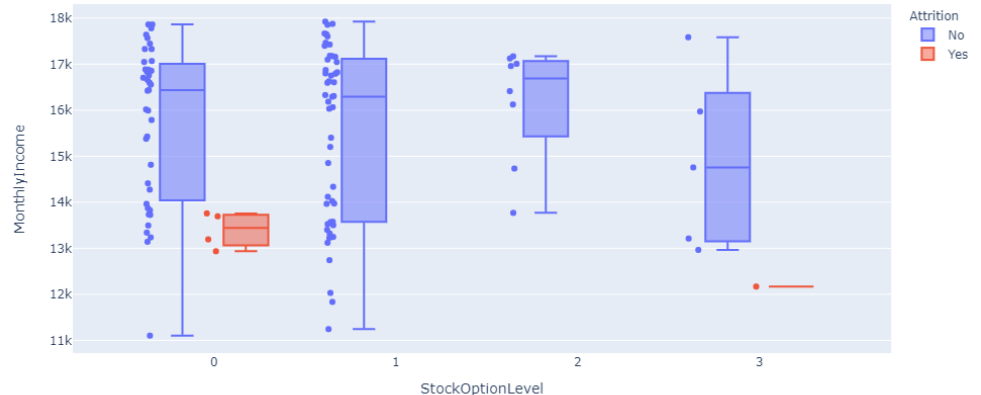
- Even though they have nearly the same median monthly income, it turns out that **employees here with attrition have a lower median stock option level** compared to those not attriting.

Job Level 2 and Stock Options



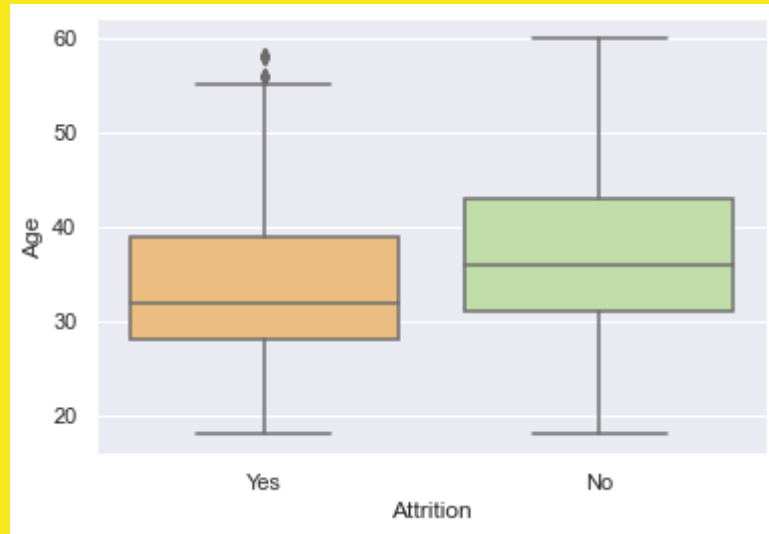
## Possible reason for **low attrition in Job Level 4:**

- Even though there is a huge gap in median monthly income, most of the employees have a **high monthly income** that may **compensate** for a lower stock option level. The red, box for those **very few employees with attrition** have **no stocks** and a much **low monthly income** compared to the rest.



# Attrition and Age

There are more younger people with attrition

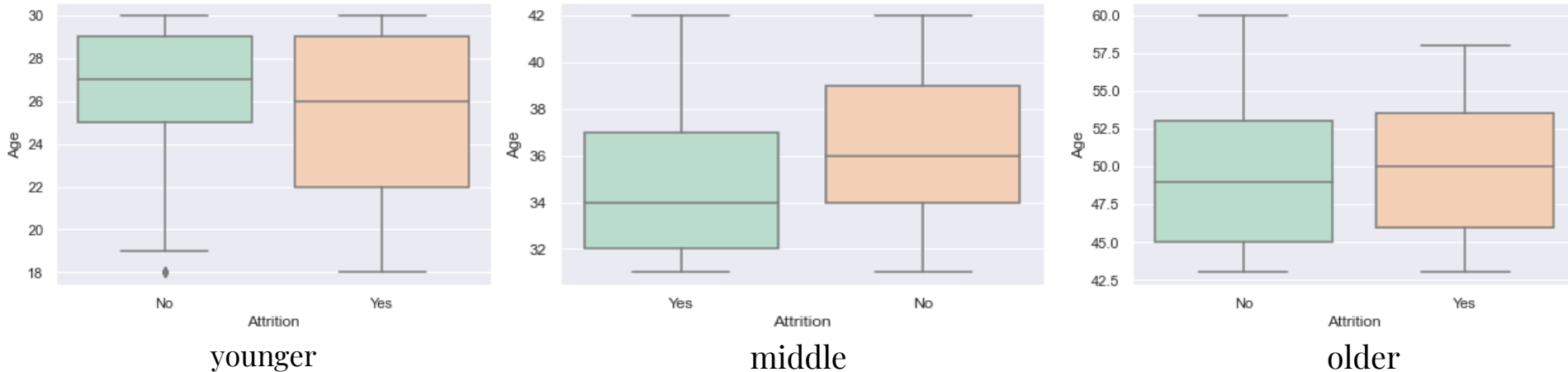


Attrition has a -0.159205 correlation with Age

Age was split into 3 categories:

- **Younger** -> aged between 18 and 30, inclusive. There is a **0.259** probability of attiring
- **Middle** -> aged between 31 and 42 inclusive. There is a **0.132** probability of attiring
- **Older** -> aged between 43 and 60 inclusive. There is a **0.116** probability of attiring

This grouping aided us in understanding from which age group are most attired employees found. These are the graphs of age against attrition:

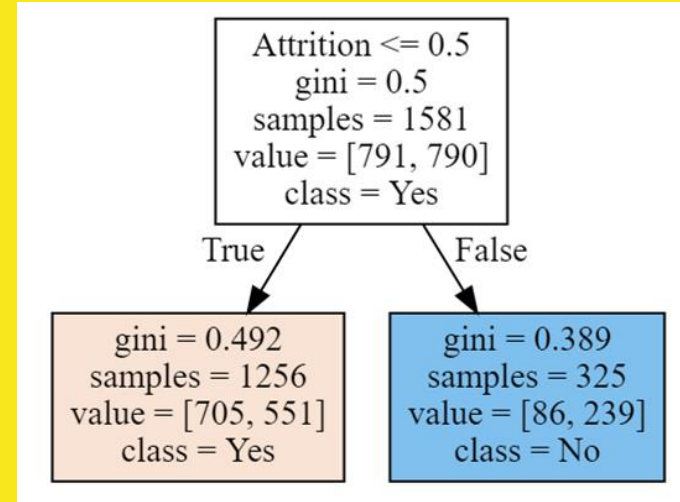


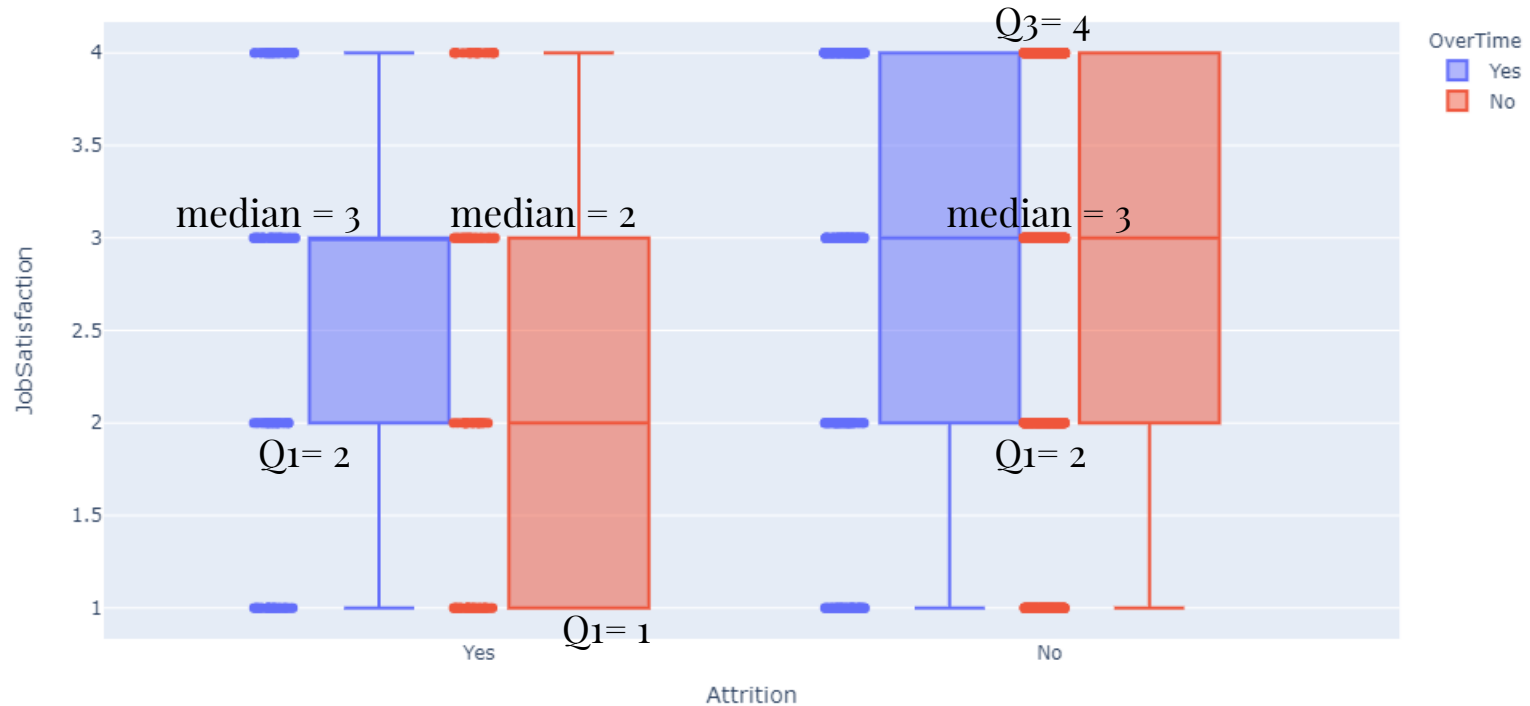
There is a higher probability for the younger group aged 18 to 30 to attire than other age groups. Employees may attire to look for better job opportunities outside the company while they are young.

# Attrition and Overtime

Employees with attrition are more likely those that do not have overtime

Attrition has a 0.246118 correlation with Overtime

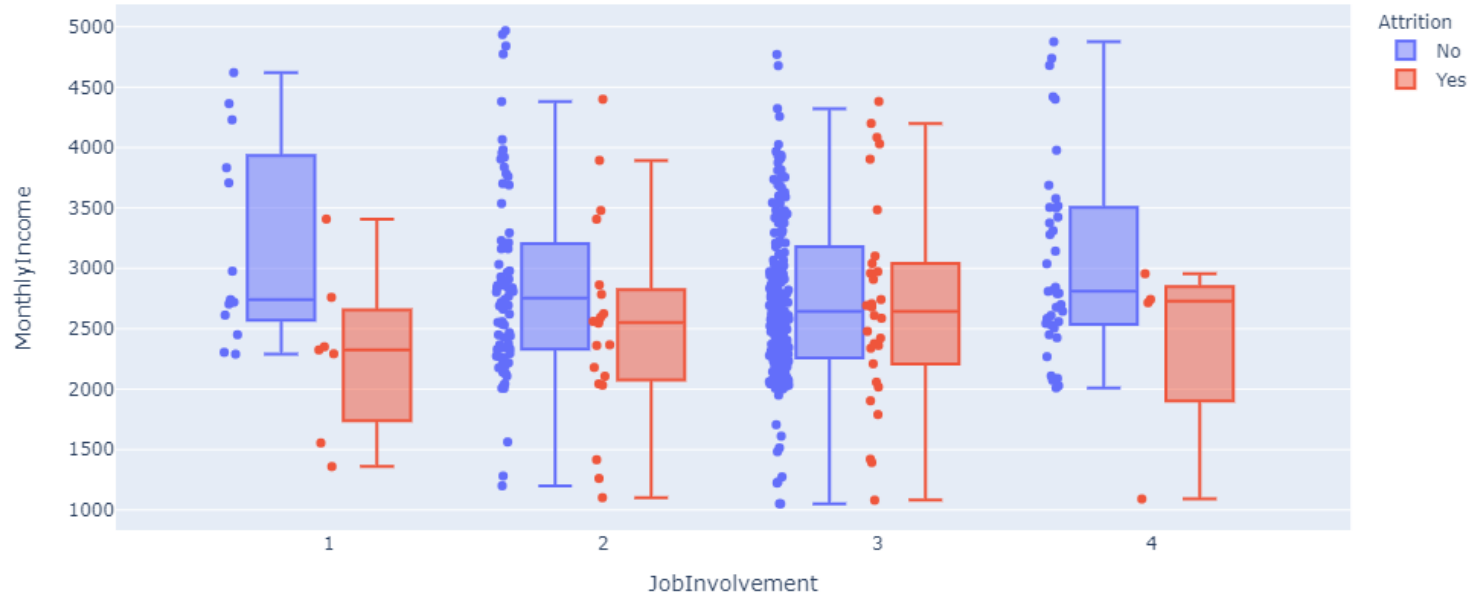




Employees without attrition, whether working overtime or not have a median job satisfaction at 3.

However, **attiring** employees **not working overtime** are actually **less satisfied** with their job than those working overtime.

## Job Level 1 Employees Not Working Overtime



In this box plot, we can see how **job level 1 employees with attrition and no overtime** have a much **lower median monthly income for every job involvement value**.

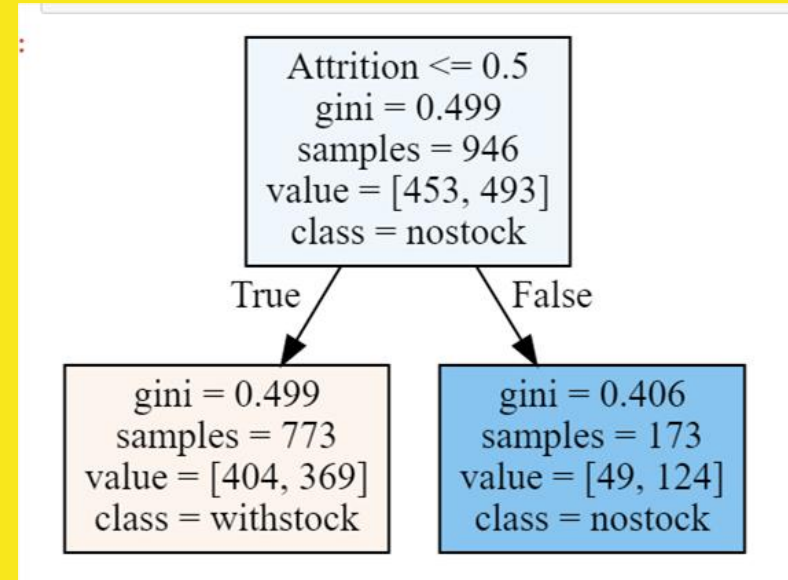
This could mean that attriting employees feel **unfair** because they are being **paid less than people in the same job level** who are also **not working overtime** and with the **same job involvement**.

*This trend is observed for all the other job levels in our Jupyter notebook*

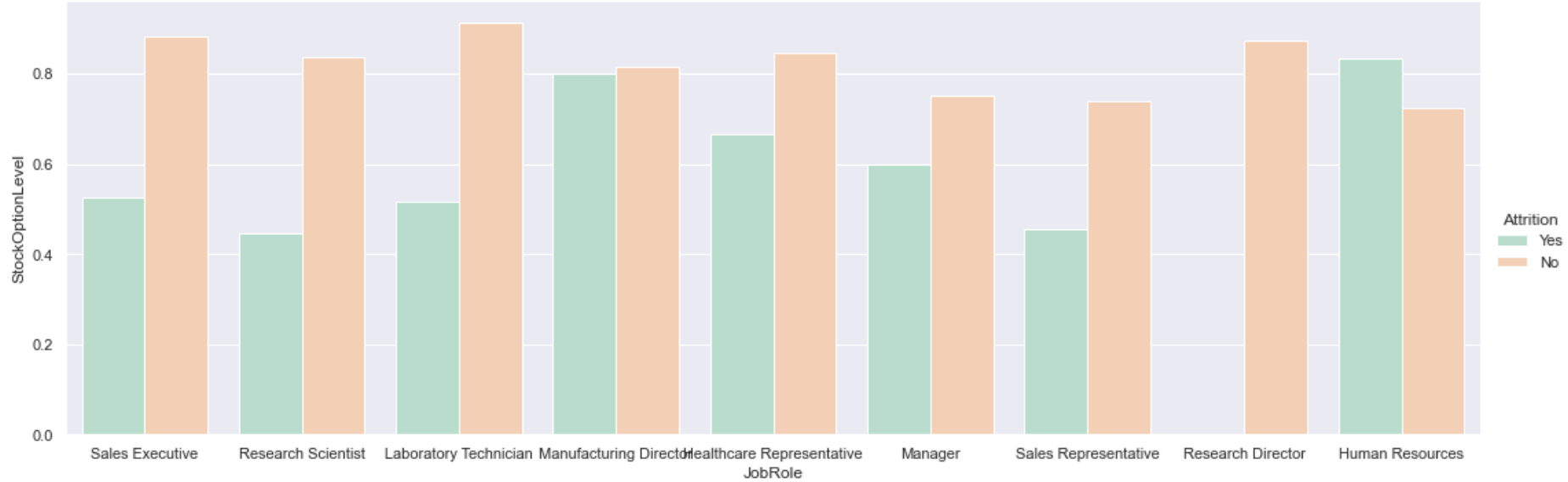


# Attrition and Stock Option Level

Employees with attrition are  
most likely those without any  
stock option



Attrition has a -0.137145 correlation with Stock  
Option Level

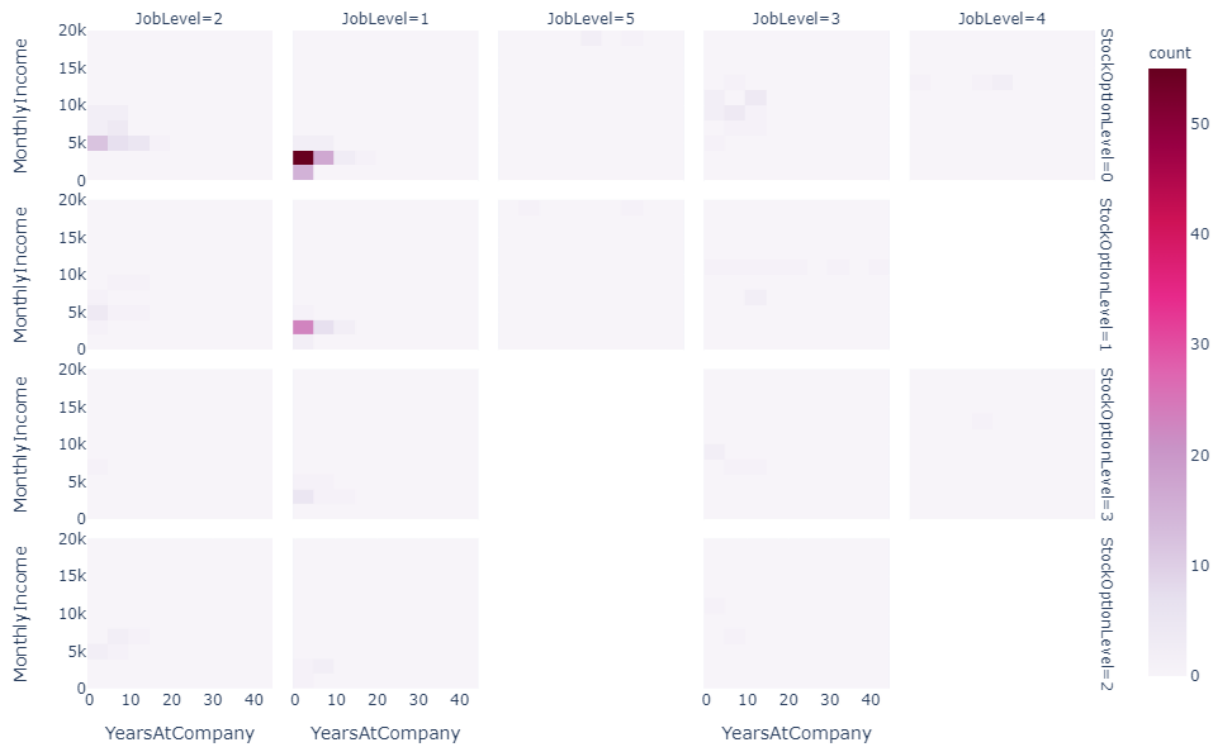


For every job role except Human Resources, there seems to be a trend where employees with attrition are those that have a lower stock level compared to those without attrition.

Employees might be **attiring** because they have a **lower stock option level** than other employees of the **same role**.

# PUTTING IT ALL TOGETHER

Density Heatmap of Employees with Attrition



We used Plotly to produce this density heatmap.

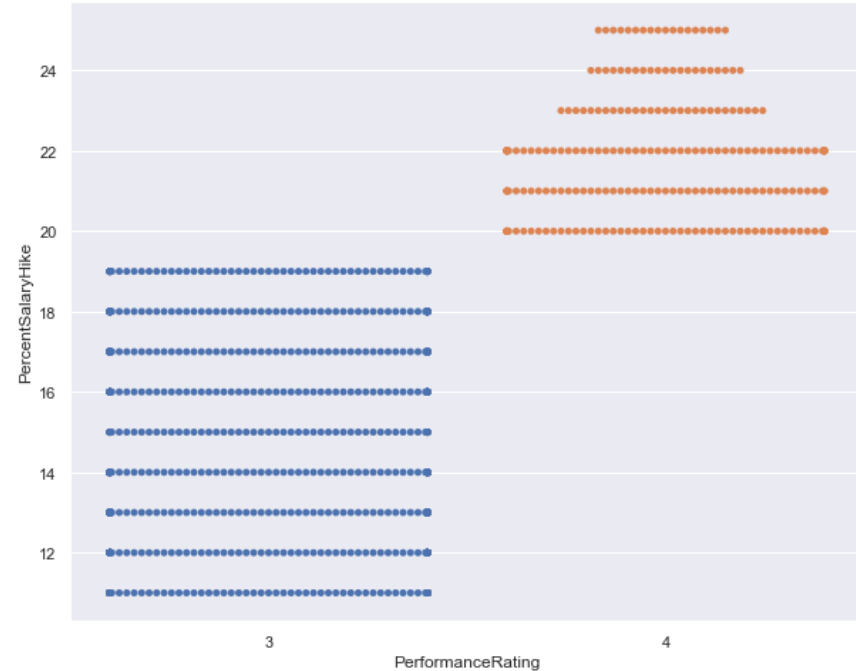
Most employees with attrition have a low Job Level, low Years at Company, low Stock Option level, and low Monthly Income

# Performance Rating of Employees

---

# WHAT IS PERFORMANCE RATING? WHAT ARE WE PLANNING TO DO WITH IT?

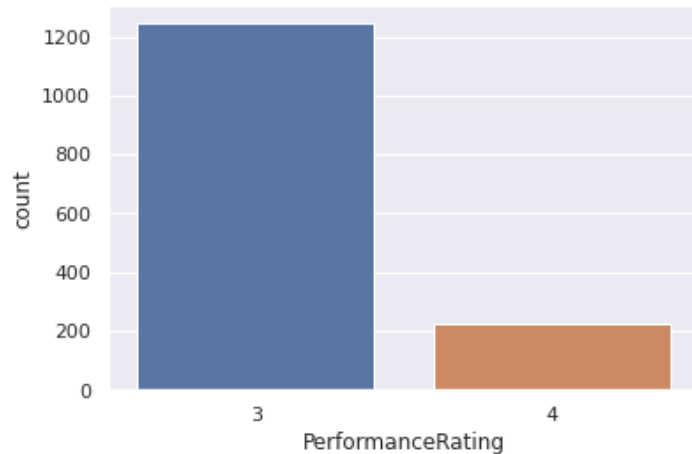
- Performance Rating is a multi-class variable which partitions the employees into 4 groups.
  - a. Performance Rating of 1
  - b. Performance Rating of 2
  - c. Performance Rating of 3
  - d. Performance Rating of 4
- However we found that the **employees only have either a performance rating of 3 or 4**
- Thus, we took performance rating as a **binary-class variable**
- Our assumption is supported by the adjacent graph. Which clearly shows, company provides **higher increase in salary to employees with Performance Rating of 4.**
- We will use **the predicted values of PerformanceRating to predict the values of PercentSalaryHike of respective employees.**



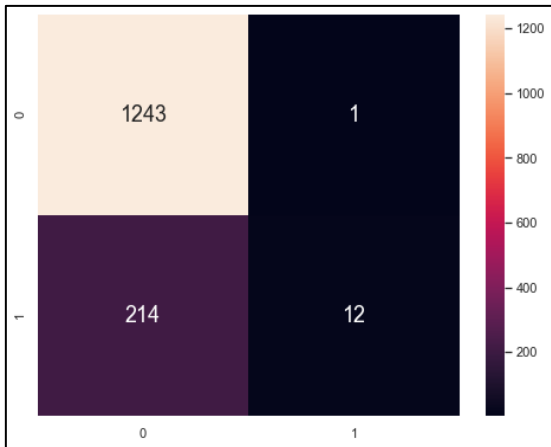
# DECISION TREE

Performance Rating was highly imbalanced so we upsampled the data using “upsample” from sklearn.utils

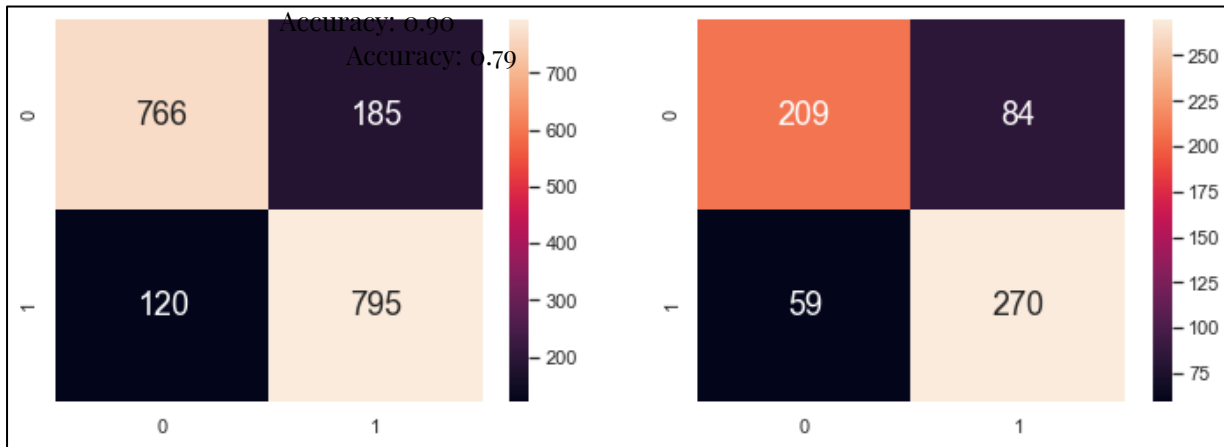
After balancing the data we ran GridSearchCV for tuning max\_depth (found it to be 10) and



Before Balancing  
Accuracy: 0.85



After Balancing Train  
Balancing Test



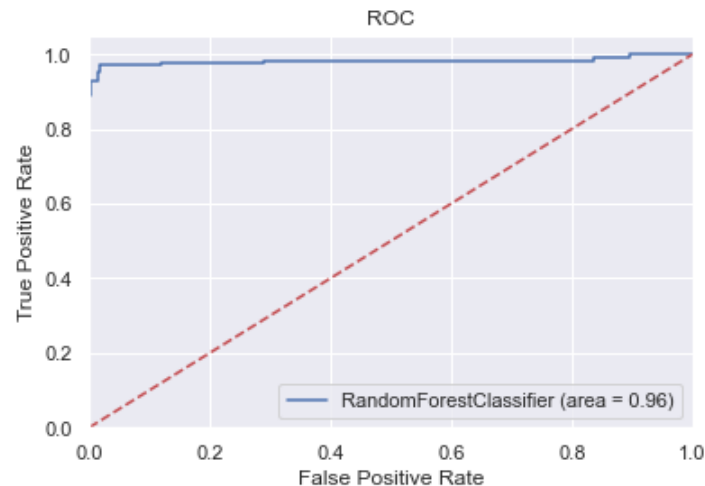
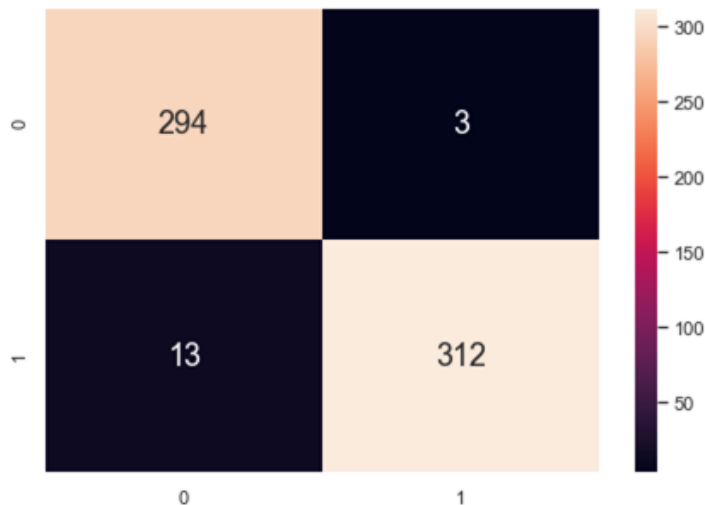
# RANDOM FOREST & VARIABLE IMPORTANCE

We used GridSearchCV to find the ideal number of trees and depth for our random forest.

```
RandomForestClassifier(max_depth=9, n_estimators=150)
```

After Balancing Test

Accuracy: 0.95

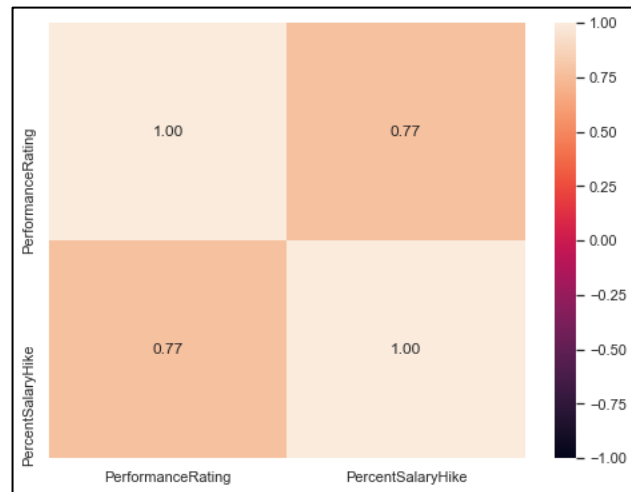


# PREDICTING PERCENT SALARY HIKE USING PREDICTED VALUES OF PERFORMANCE RATING

We know that PercentSalaryHike and PerformanceRating are perfectly correlated in terms of classification, whereas linearly has a correlation of 0.77.

So, we can treat PercentSalaryHike in two ways :

- As a continuous variable and perform a regression model.
  - RandomForestRegressor
- As a multi-class variable and perform a multi-classification model.
  - RandomForestClassifier





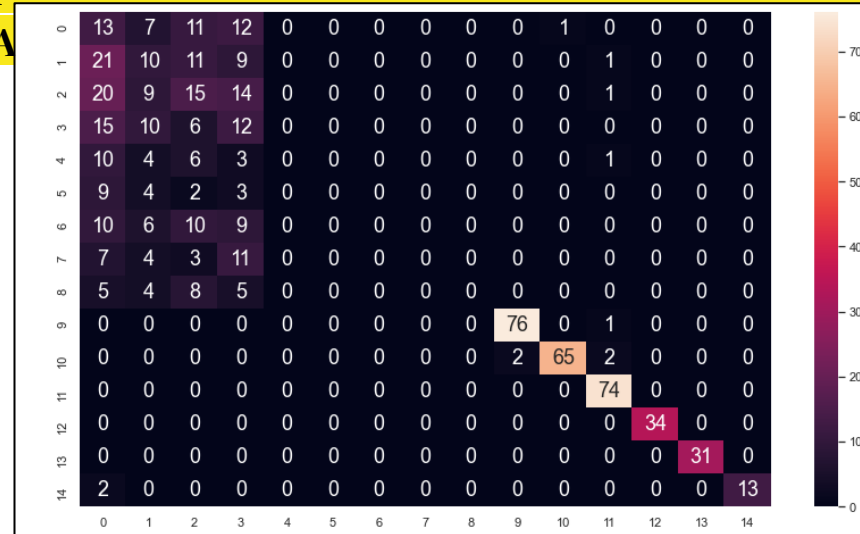
# RandomForestClassifier

RandomForestRegressor

RandomForestClassifier

Accuracy Test: 0.841

Classification A



So comparing the accuracy of the models we decided to go with the **regression model** to predict **PercentSalaryHike**

“Research indicates that workers have three prime needs:  
Interesting work, recognition for doing a good job, and  
being let in on things that are going on in the company.”

---

- Zig Ziglar

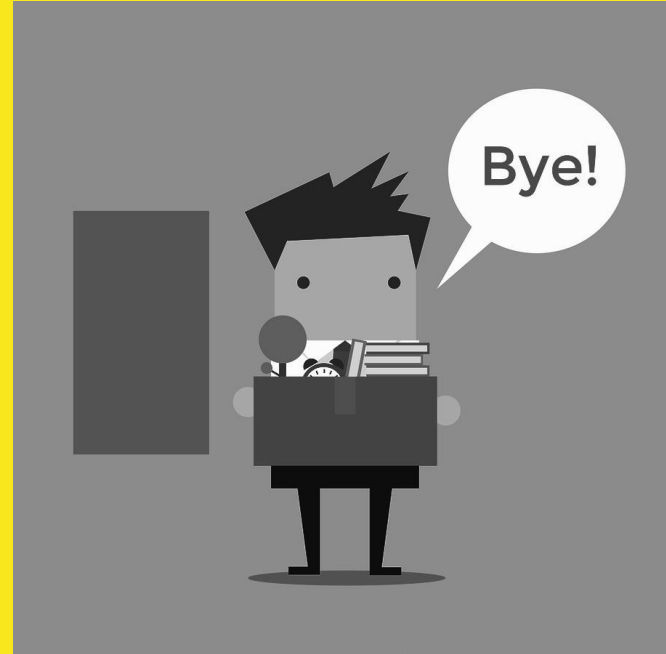
# CONCLUSION

To have a low attrition rate, it is important for companies to:

1. Fairly pay employees in same job level, same job involvement, same job role with almost equal monthly income and stocks.
2. Offer stock options to more employees especially new recruits as an incentive.
3. Increase allowance for employees to work overtime, and make sure that those working overtime are being paid more than those who are not.
4. Also from our quote, companies should remember to appreciate their employees and work on the harmonic balance of their company.

# THANK YOU

From: DS2 Group 7



# JOB DISTRIBUTION

*From the things that are in this presentation*

**Data Cleaning :** Angelin

**Exploratory Data Analysis:**

Shivangi (numeric), Angelin (categorical), Pratham (categorical)

**Linear Regression:** Shivangi and Pratham

**Random Forest & ROC/AUC Curve (Attrition):**

Angelin

**Random Forest Regressor & Classifier (Performance Rating):**

Pratham