Employee Attrition Analysis

Group 7

Introduction

- Employees matter!

Employees are the most valuable assets of a company. Extraordinary and reliable employees can create huge wealth for the company. However, nowadays, managers are perplexed by high employee attrition rate and they wonder what the attrition rate of their companies would be and in which direction they can work hard to improve employees' loyalty.

- Predictive model of employee attrition

Thus, in this project, we will use IBM HR Analytics Employee Attrition & Performance dataset to construct a predictive model of employee attrition. Also, we will find what factors influence employee attrition most and what managers can do to improve employees' loyalty and satisfaction.

Project Objectives

some questions this project will answer :

1. What would attrition probability of employees be given features of employees?

2. What factors influence employee attrition most and what can managers do to improve employees' satisfaction and loyalty?

3. Does there exist gender inequality issues in the company?

Data Source and Description

- Data Source
  IBM HR Analytics Employee Attrition & Performance dataset is from Kaggle.

- Data Description
  The data have 1470 observations and 35 variables. These variables are divided into three categories:
  - Personal Information: Age, Education, Education Field, Gender, Over 18, Marital Status.
  - Job Information: Attrition, Business Travel, Daily Rate, Department, Distance from Home, Employee Count, Employee Number, Hourly Rate, Job Environment,Job Level, Job Role, Monthly Income, Monthly Rate, Num Companies worked, Over Time, Percent Salary Hike, Standard Hours,Stock option Level, Total Working Years, Training Times Last Year, Years at Company, Years in Current Role, Years Since Last Promotion, Years With Current Manager.
  - Degree of Satisfaction: Environment Satisfaction, Job Involvement, Job Satisfaction, Performance Rating, Relationship Satisfaction, Work Life Balance.

Exploratory Data Analysis

- Age and Education Analysis
- The Impact of Income towards Attrition
- Working Environment
- Gender Analysis

General Information

- our data frame consists of 1470 observations of 35 variables

- only 2 kinds of data types:

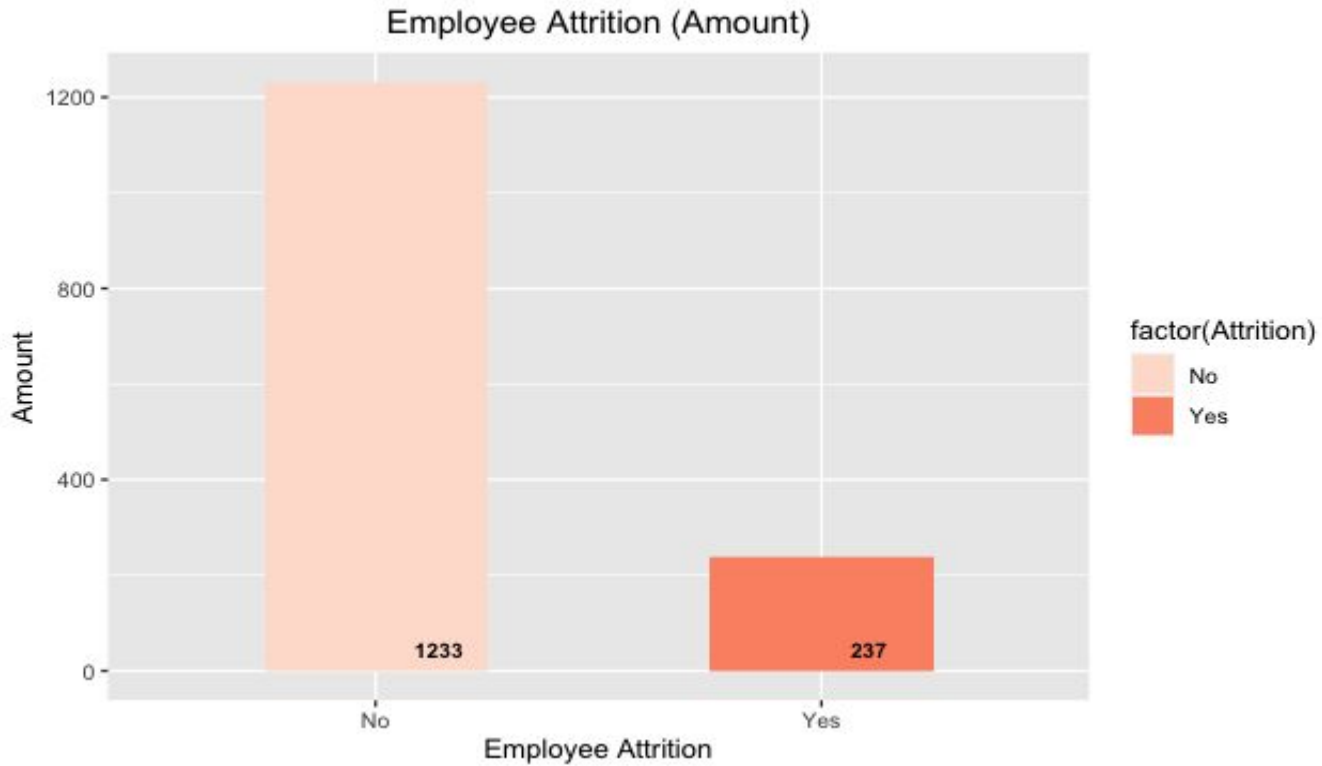
1) Integers
2) factors

- no missing data in our dataset

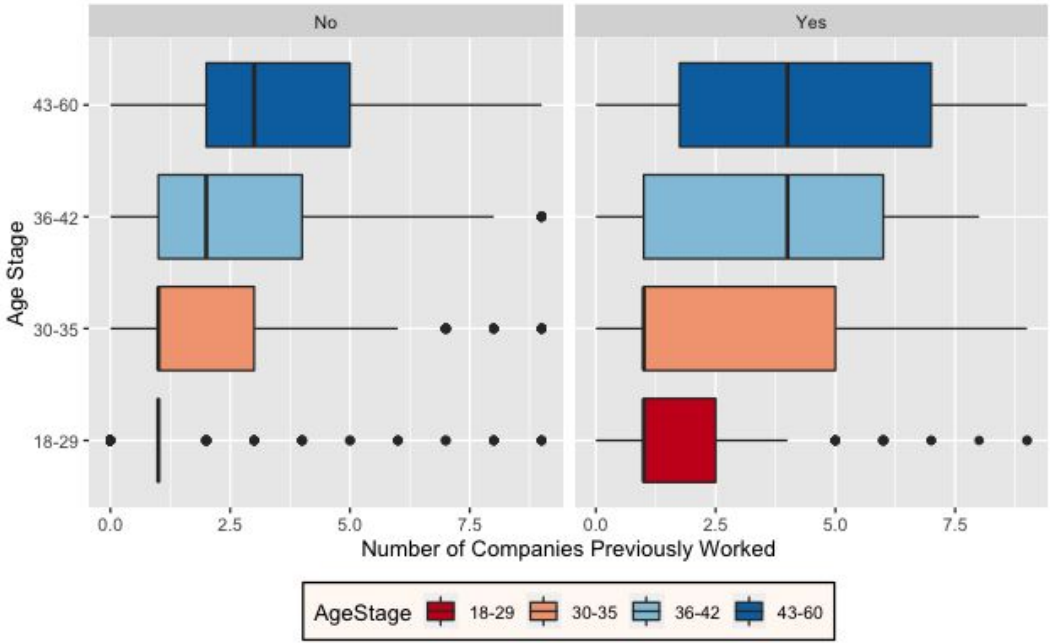
- All the variables can be categorized into 2 types:

1) categorical variable
2) numerical variable

General Information



Number of Companies Worked by Age Stage



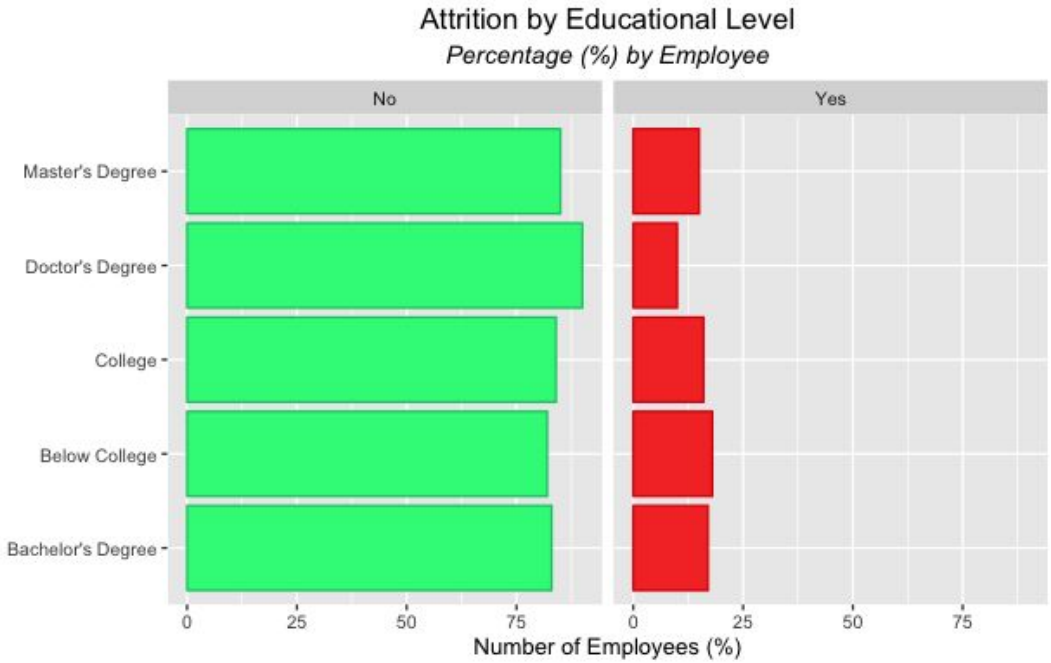
	No	Yes
43-60	335	44
36-42	329	33
30-35	334	69
18-29	235	91

H0: No association between age and attrition choice
H1: There is association

Pearson's Chi-squared test

data: Agedata
X-squared = 52.665, df = 3, p-value = 2.161e-11

Attrition by Educational Level

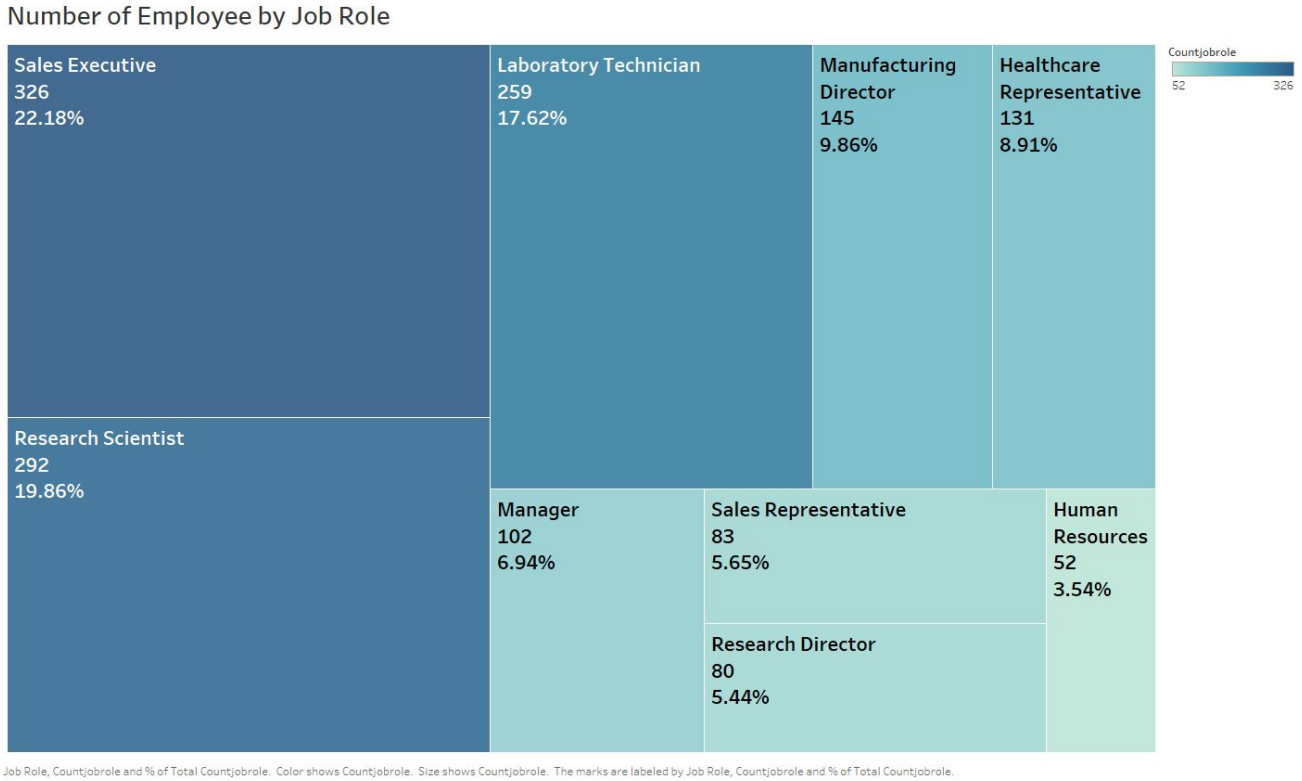


	Under Master's	Above Master's
Yes	174	63
No	850	383

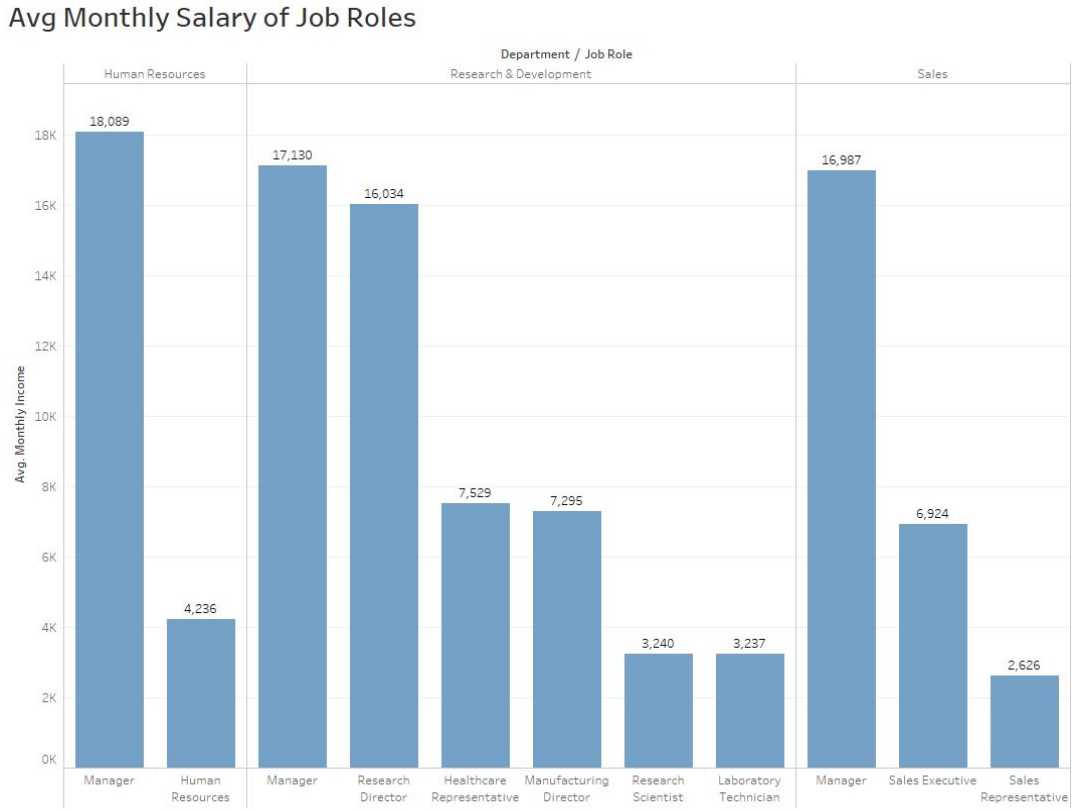
H0: p1 ≤ p2
H1: p1 > p2

data: x out of n
X-squared = 1.6819, df = 1, p-value = 0.09734
alternative hypothesis: greater
95 percent confidence interval: -0.00623751 1.00000000
sample estimates:
prop 1 prop 2
0.1699219 0.1412556

Overview of Employee Structure

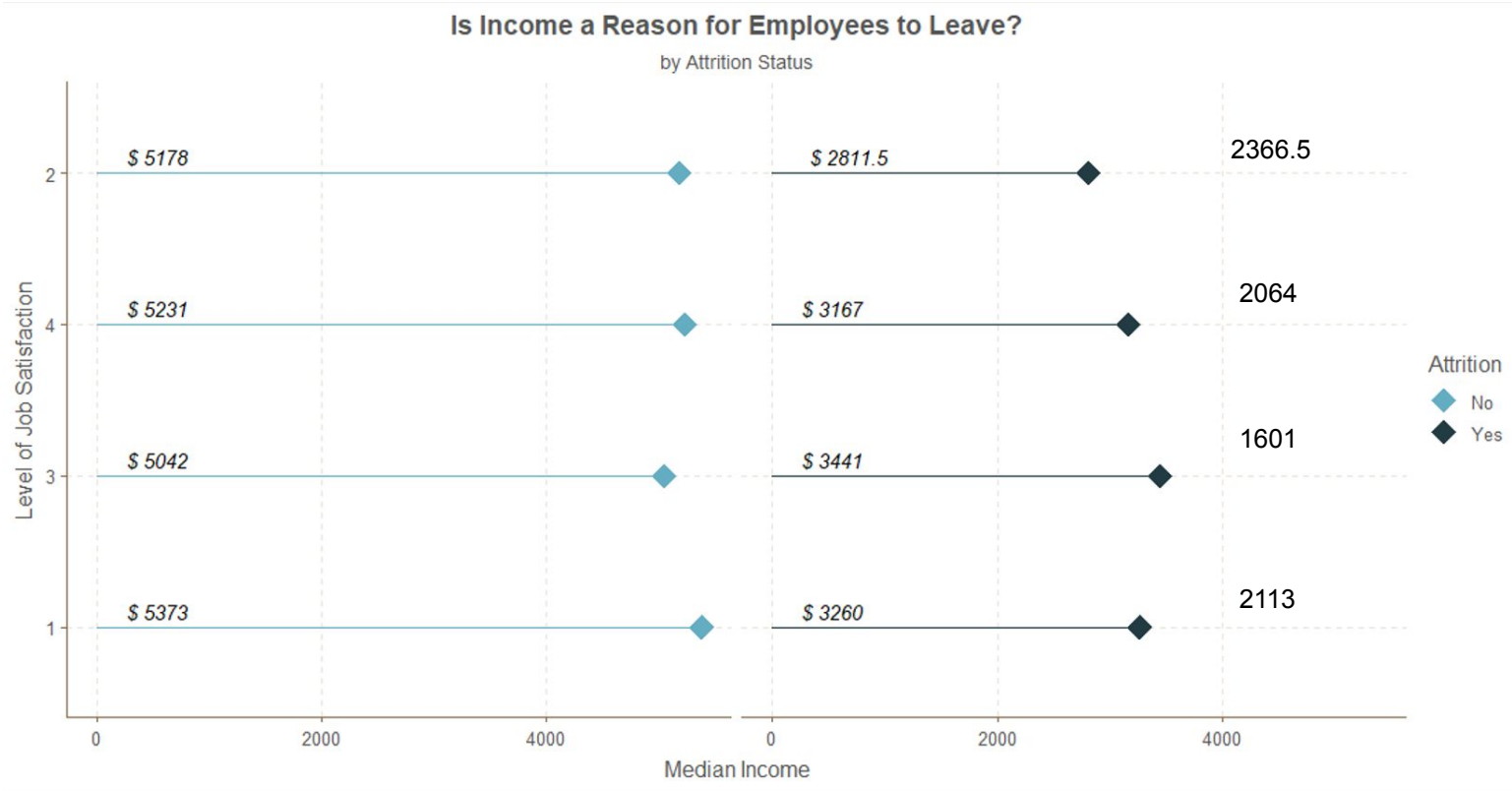


Average Monthly Salary by Job Role



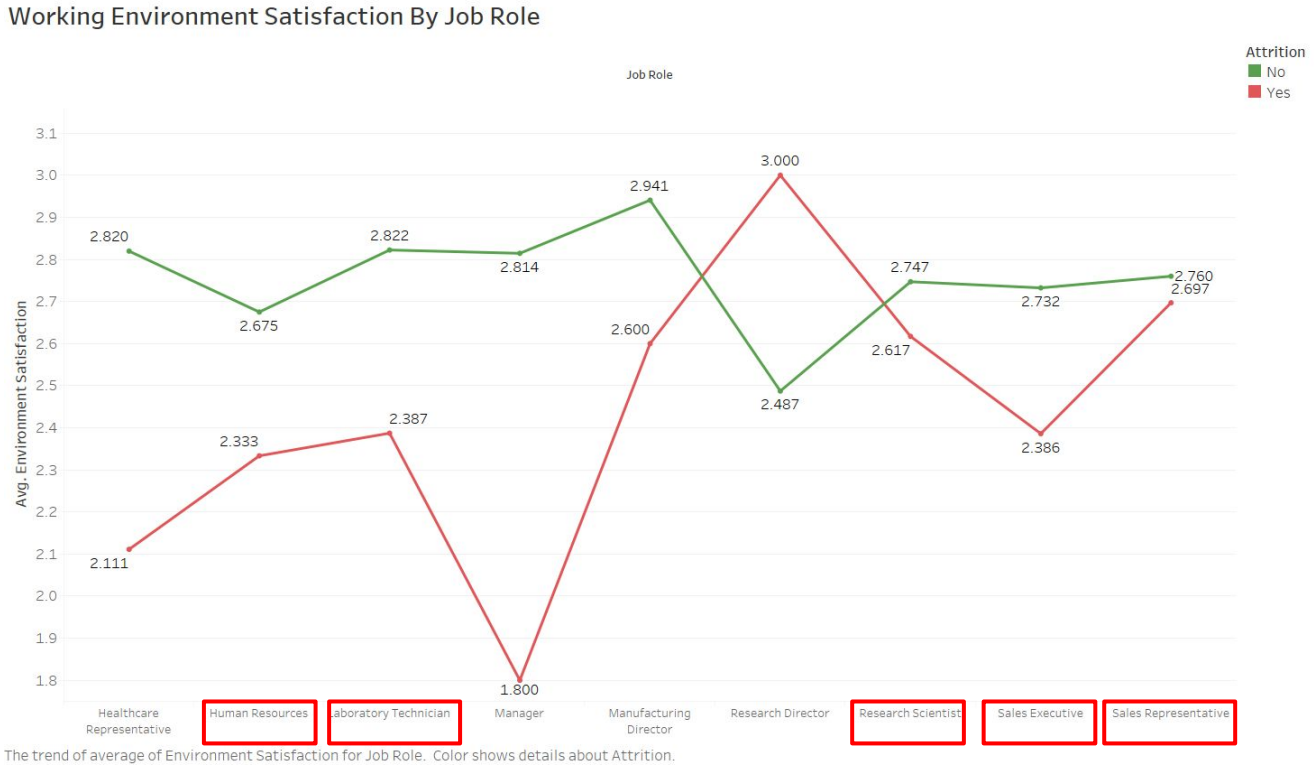
Average of Monthly Income for each Job Role broken down by Department. The marks are labeled by average of Monthly Income.

Determining Satisfaction by Income

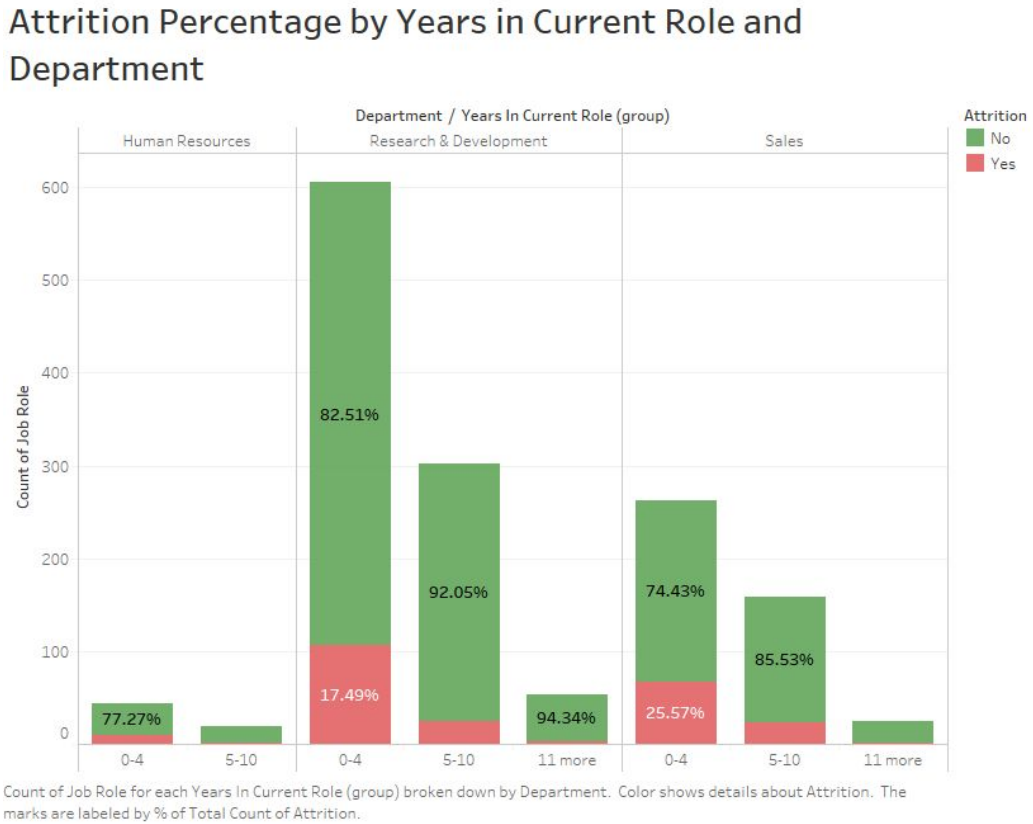


Attrition Percentage by Job Role

Should the Company Improve Working Environment?



Should the Company Offer Promotions?

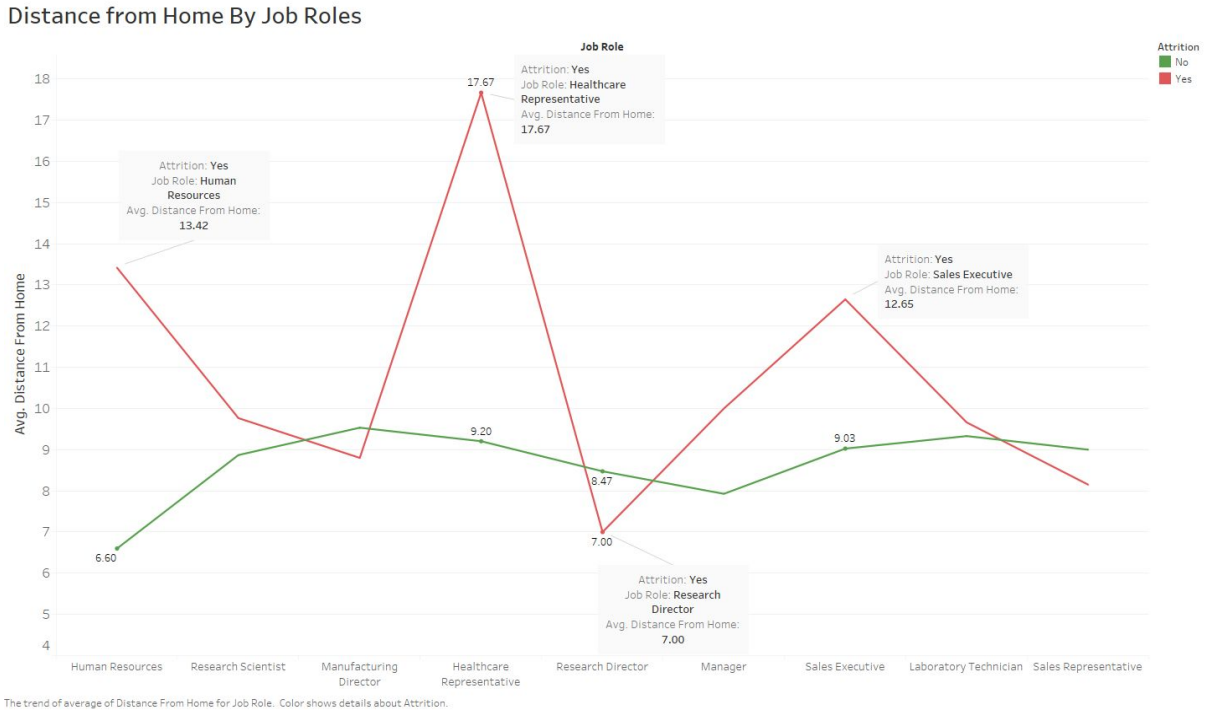


Should the Company Offer More Training Sessions?

Should Company Offer Housing Reimbursement?

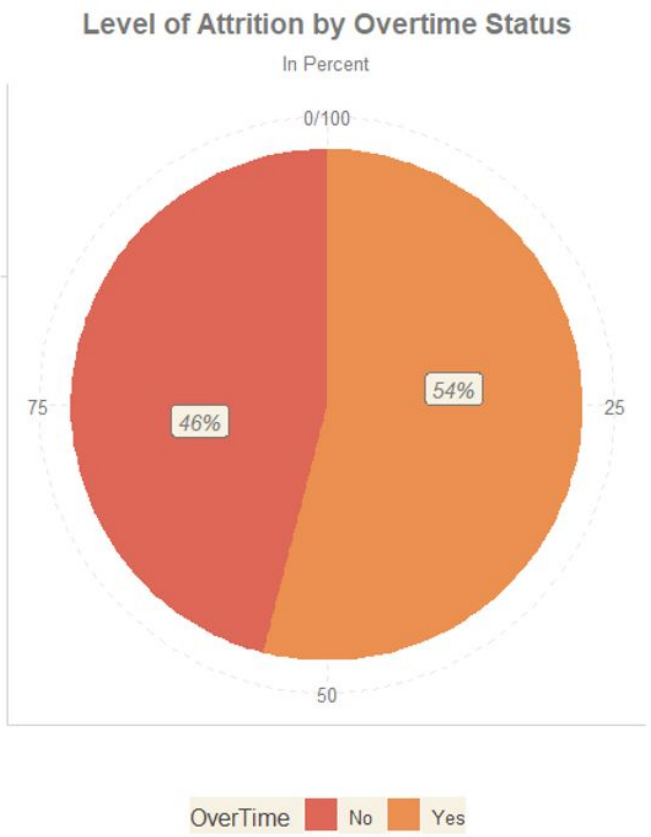
Job Role	
Healthcare Representative	9.786
Human Resources	8.173
Laboratory Technician	9.409
Manager	8.029
Manufacturing Director	9.483
Research Director	8.438
Research Scientist	9.014
Sales Executive	9.660
Sales Representative	8.663

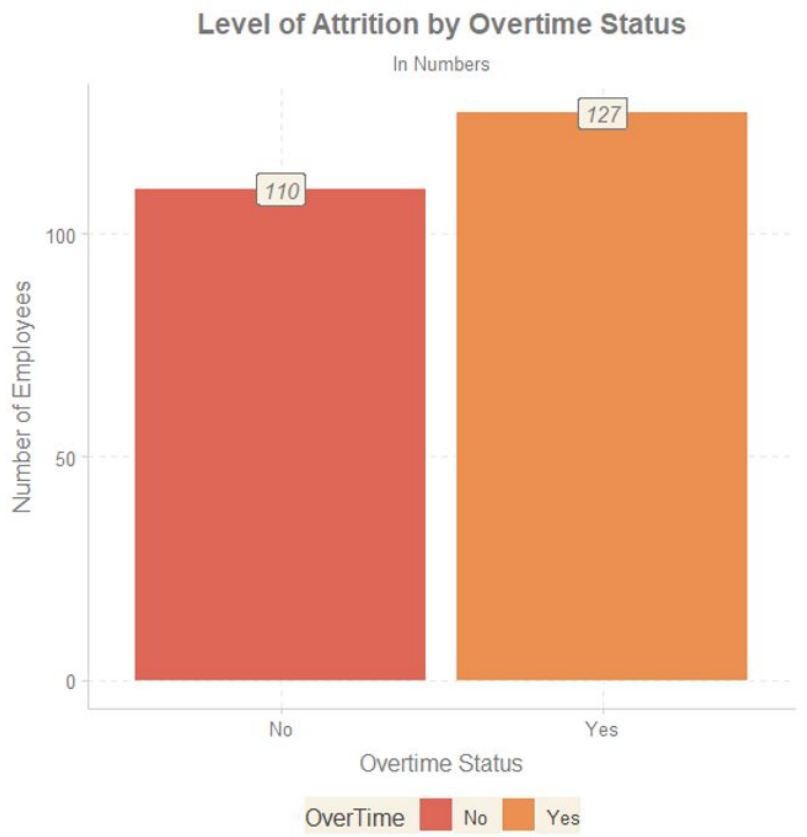
Average of Distance From Home broken down by Job Role.



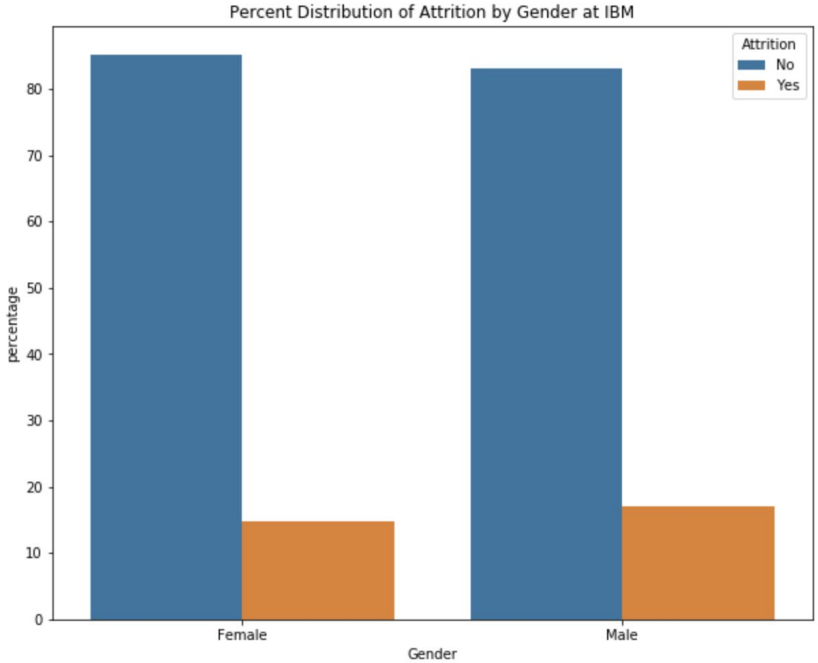
The trend of average of Distance From Home for Job Role. Color shows details about Attrition.

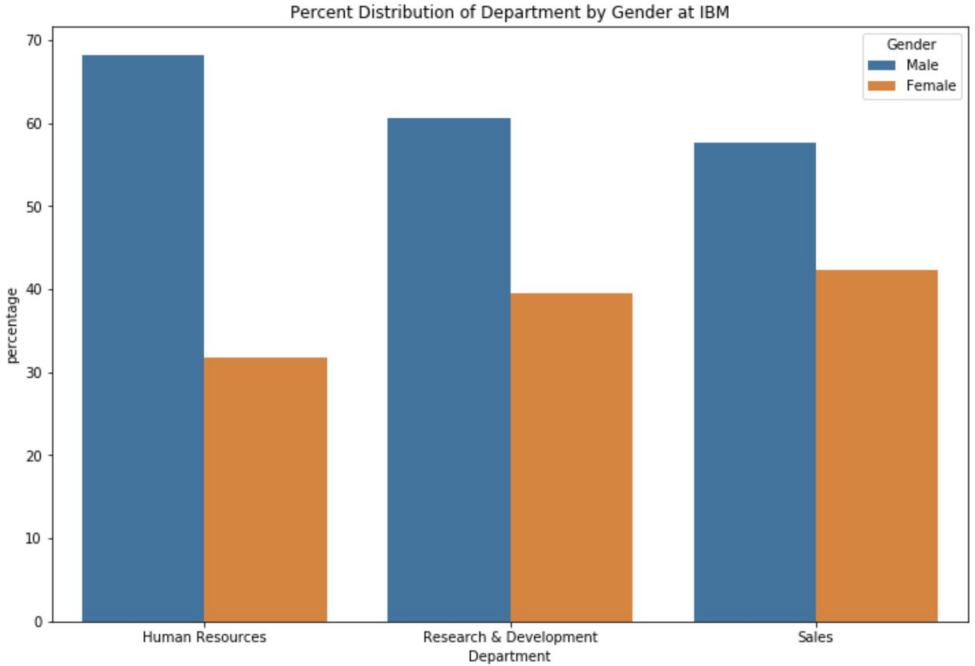
Should Company value work-life balance?



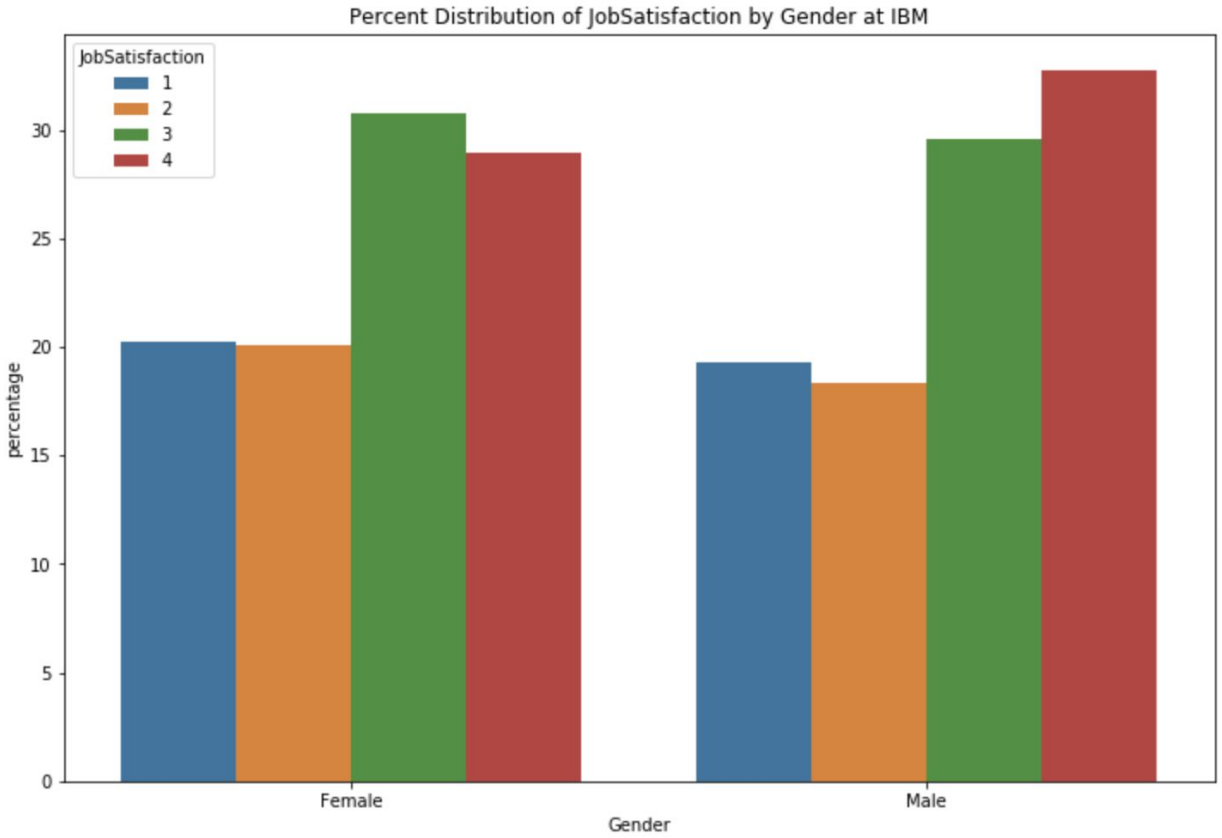


Attrition Distribution by Gender



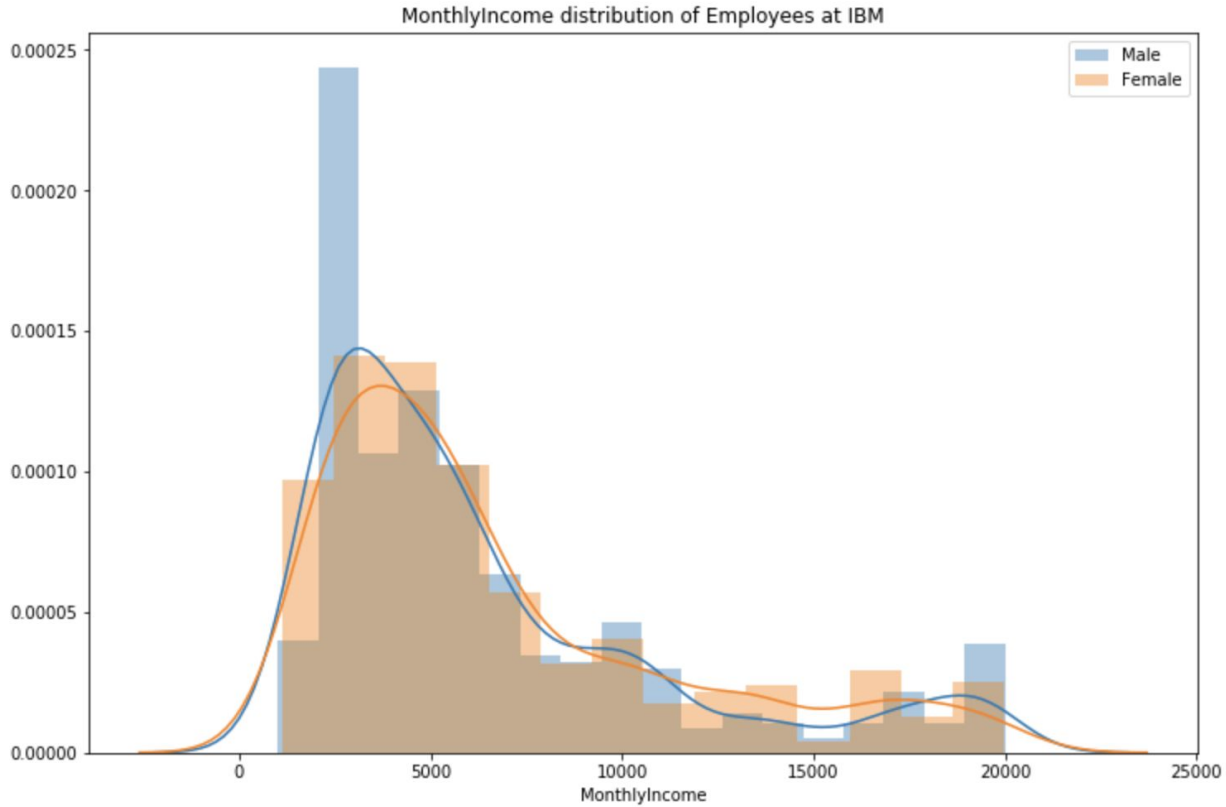


Job Satisfaction Distribution by Gender



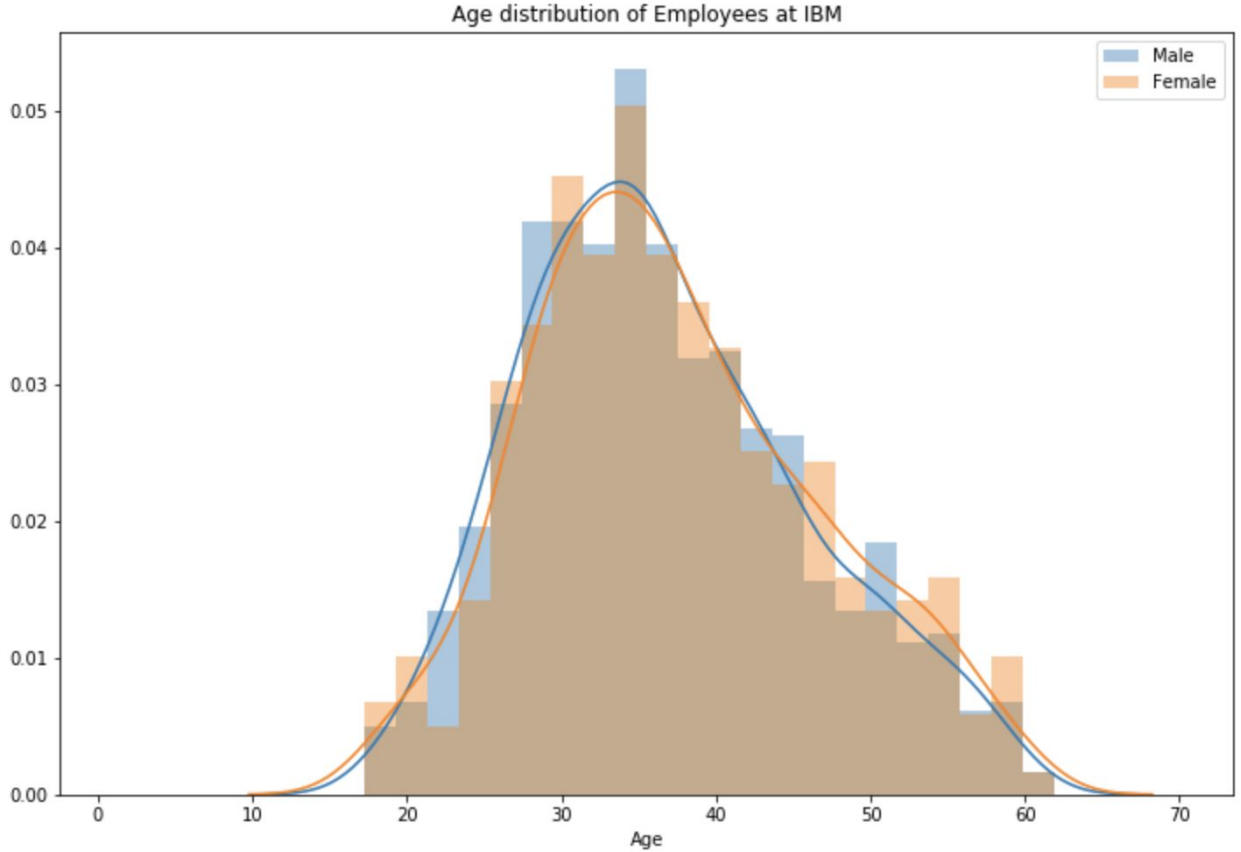
Gender	JobSatisfaction	percentage
Female	3	30.782313
Female	4	28.911565
Female	1	20.238095
Female	2	20.068027
Male	4	32.766440
Male	3	29.591837
Male	1	19.274376
Male	2	18.367347

Monthly Income Distribution by Gender



Gender	count	mean	std
Female	588.0	6686.566327	4695.608507
Male	882.0	6380.507937	4714.856577

Age Distribution by Gender



Gender	count	mean	std
Female	588.0	37.329932	9.266083
Male	882.0	36.653061	9.042329

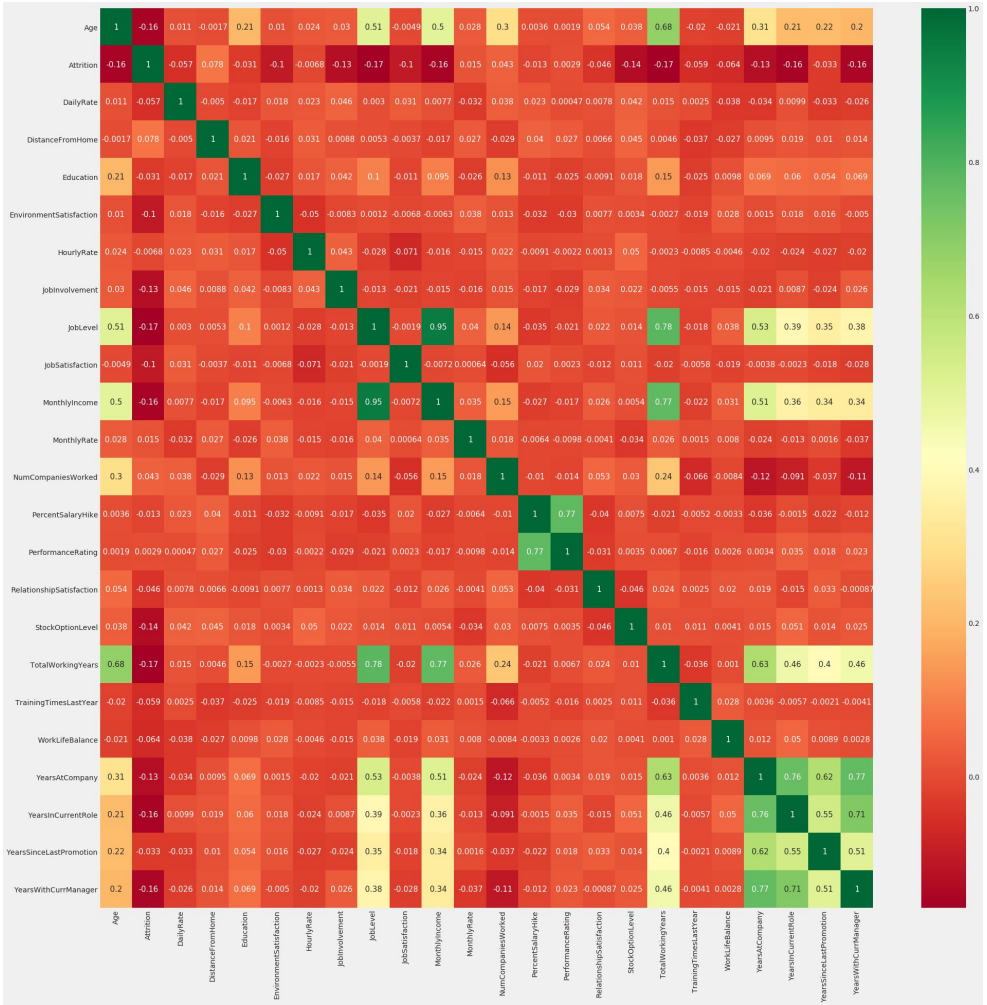
Applying Machine Learning Algorithms

- Correlation Matrix
- Data Processing
- Train different machine learning models
- Best Machine Learning model & Feature Importance
- Conclusion

Correlation Matrix

First, we use *heatmap* to check the correlation matrix of different numerical variables.

We can clearly see that the correlation between most numerical variables is not strong.



Data Processing: Dummies

First of all, we noticed that we have many categorical variables. These variables are also important factors that affect Attrition.

In order to apply these factors to machine model algorithms other than tree models, we transform categorical data into dummies.

Data Processing: Train-test Split

- We split all our data into the training set and the test set at a ratio of 2:8.
- Checking that both the training and testing sets have the same label proportions.The attrition ratio of the two sets is the same

A tibble: 2 × 3

Attrition	n	pct
<fct>	<int>	<dbl>
No	987	0.84
Yes	190	0.16

A tibble: 2 × 3

Attrition	n	pct
<fct>	<int>	<dbl>
No	246	0.84
Yes	47	0.16

Applying Machine Learning Algorithms

What defines success?

```
===============TRAIN=================
Staying Rate: 83.87%
Leaving Rate: 16.13%
===============TEST==================
Staying Rate: 83.90%
Leaving Rate: 16.10%
```

Apply four machine learning models

	Test accuracy	Train accuracy	Test Precision	Recall Score	F1 score	Confusion Matrix
Logistic Regression	85.26%	92.91%	56.00%	39.44%	46.28%	[[348 22] [43 28]]
Random Forest Classifier	83.90%	100%	50.00%	8.45%	14.46%	[[364 6] [65 6]]
Decision Trees Classifier	85.49%	94.07%	61.29%	26.76%	37.25%	[[358 12] [52 19]]
XGBoost Classifier	85.48%	88.82%	46.15%	29.51%	36.00%	[[359 21] [43 18]]

BEST model Decision Tree

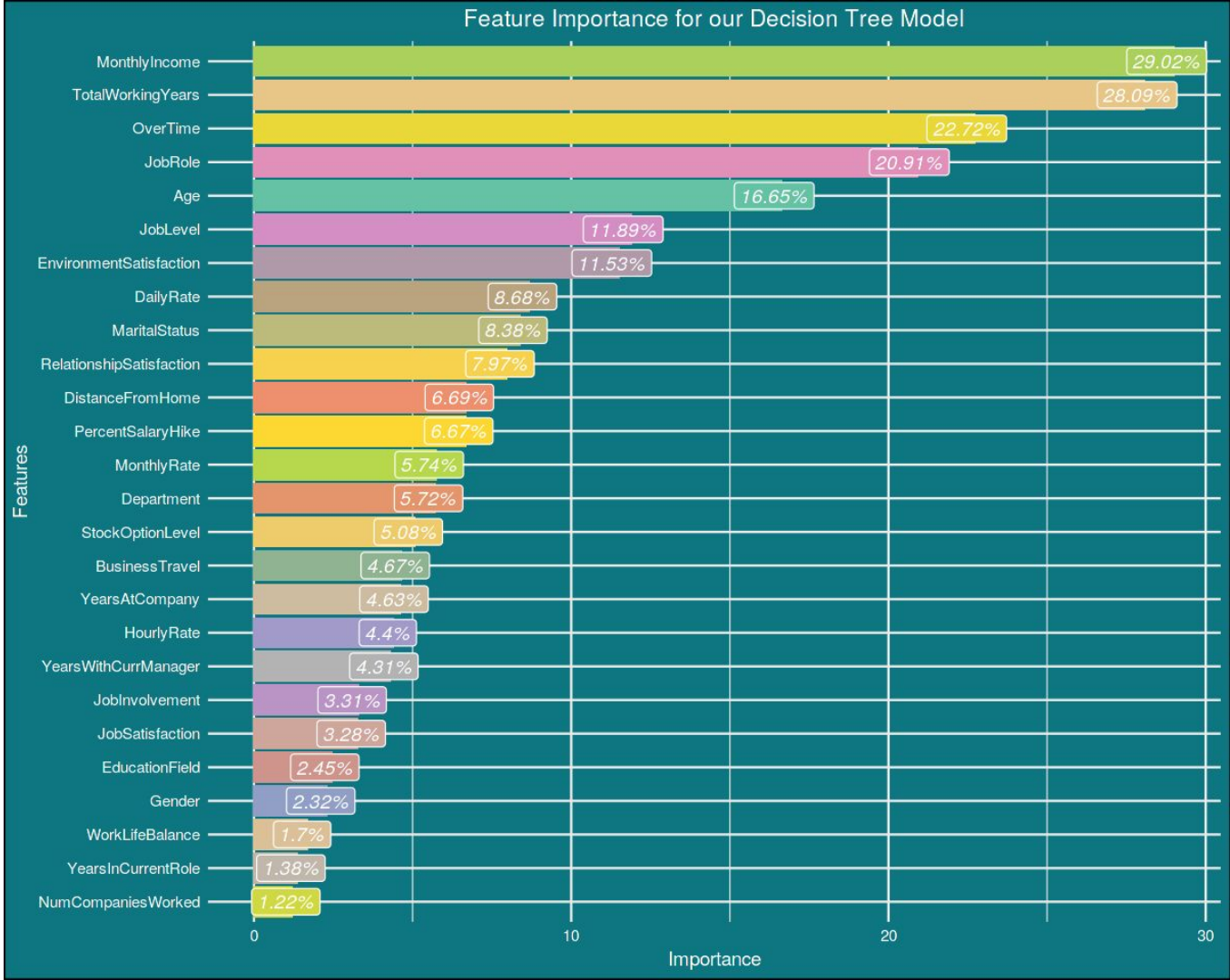
Monthly Income

Total Working Years

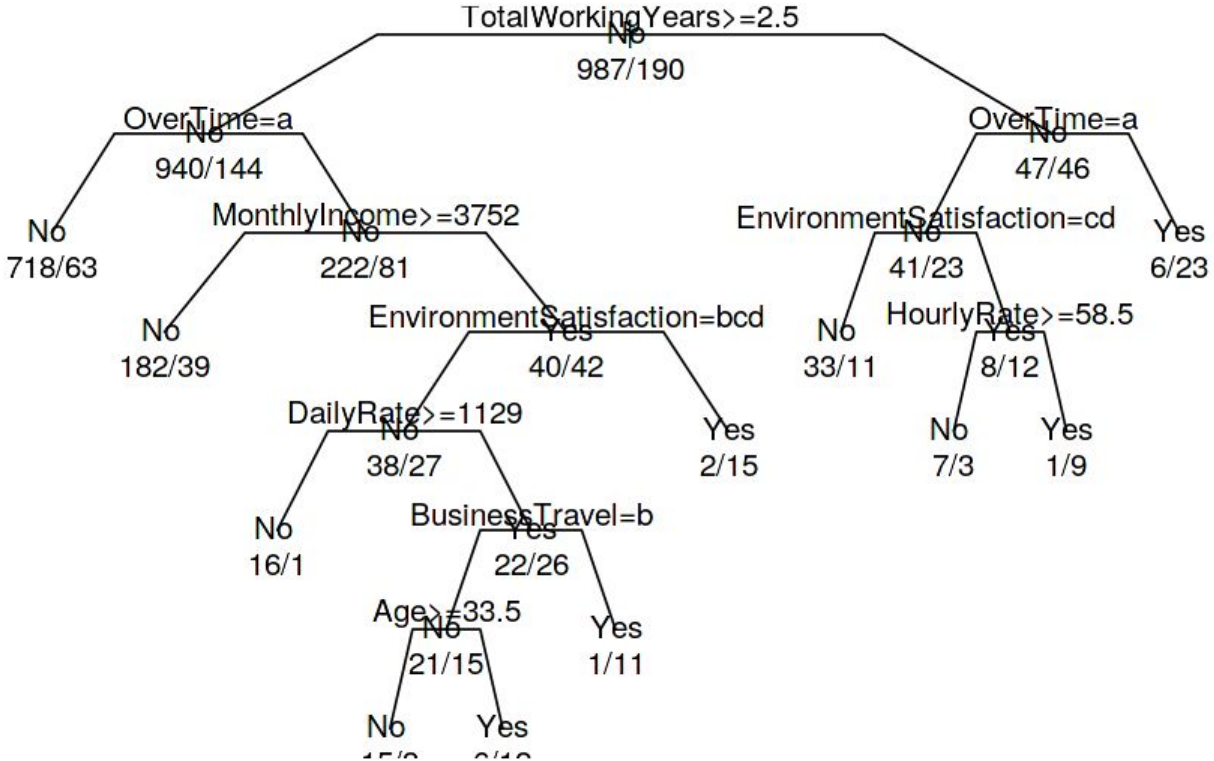
OverTime

Job Role

Age



Classification process



Conclusion

The Decision Tree Model considers these six attributes as the most important:

- **Monthly income** :As expected, Income is a huge factor as why employees leave the organization in search for a better salary.
- **Total Working Years**: A portion might be retiring or looking for more challenges.
- **No Overtime**: This was a surprise, employees who don't have overtime are most likely to leave the organization. This could be that employees would like to have a higher amount of income or employees could feel that they are underused.
- **Job Role**: Employees who do not like their current Job Positions.
- **Job Level:** Employees who believe they deserve a higher job level are prone to leaving the organization.
- **Age**:This could also be expected, since people who are aiming to retire will leave the organization.