HR ANALYTICS- ATTRITION PROBLEM

SUBMITTED BY SHUBHAM BANSLA

1. The Problem Statement

In the best of world's, employee would love their jobs, like their co-workers, work hard for their employers, get paid well for their work , have ample chances for advancement, and flexible schedules so they could attend to personal or family needs when necessary and never leave.

But then there's the real world. And in the real world, employees, do leave, either because they want more money, hate the work condition, hate their co-workers, want a change, or because their spouse gets a dream job in another state.

It is however not an easy task for an HR manager to bridge the ever increasing demand and supply gap of professionals. HR manager is not only required to fulfill this responsibility, but also find the right kind of people who can keep the unique pace with the unique work patterns in industry. Adding to this is the issue of maintaining consistency in performance and keeping the motivational level high, despite the monotonous work. The toughest concern for an HR manager is however the high attrition rate.

2. The objective

There are various objective which are as follows:-

- To predict if an employee is going to leave or not
- What are the various factors involved in the process of deciding job objectives for employed people.
- To find out the similarities and differences in the decision making process for when employee satisfaction with respect to the job profile and the organization changes.
- Find how these factors for employees are influenced by demographic differences.

3. Methodology

- Through our analysis we intend to build a model which can predict if an employee is about to quit.
- We shall be looking at all variable through some plots and infer about it in our exploratory analysis
- After our exploration we shall build some features based on the variables at hand and take a call on inclusion and exclusion of few variable

4. Exploratory Data Analysis

In this Section we are going to analysis each variable or feature present in the data set along with inference about their distribution.

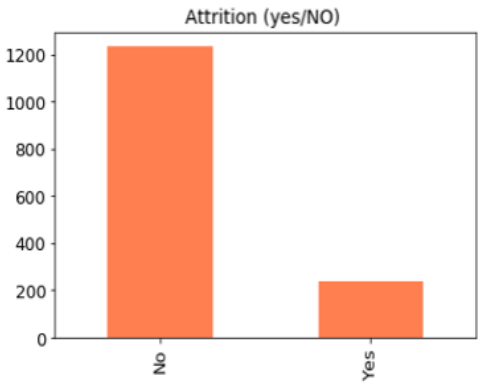
4.1 Variable and Their Types

Name Of The Variable	Type Of Variable
Age	Numerical
Attrition.	Categorical
Business Travel	Categorical
Daily Rate	Numerical
Department	Categorical
Distance from Home	Numerical
Education	Categorical
Education Field	Categorical
Employee Count	Numerical
Employee Number	Numerical
Environment Satisfaction	Categorical
Gender	Categorical
Hourly Rate	Numerical
Job Involvement	Categorical
Job Level	Categorical
Job Role	Categorical
Job Satisfaction	Categorical
Marital Status	Categorical
Monthly Income	Numerical
Monthly Rate	Numerical
Number of Companies Worked	Numerical
Over18	Categorical
Over Time	Categorical
Percent Salary Hike	Numerical
Performance Rating	Categorical

Relationship Satisfaction	Categorical
Standard Hours	Numerical
Stock Option Level	Categorical
Total Working Years	Numerical
Training Times Last Year	Numerical
Work Life Balance	Categorical
Years at Company	Numerical
Year since Current Role	Numerical
Years Since Last Promotion	Numerical
Years with Current Manager	Numerical

4.2 Distribution of the Whole Data

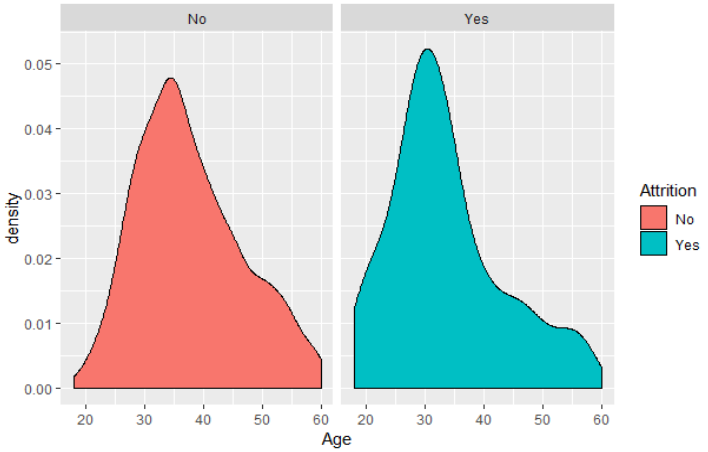
- In the data set there are 1047 rows and 35 columns.
- There is no missing value present in the data so we can easily omit the missing value step.
- Our target Variable is "Attrition" whereas other 34 variables are predictors which will predict the "Attrition either Yes or NO".
- Values which are showing **"Attrition=Yes"** are only 16% of the whole data where 84% values showing "**Attrition = No**", there may be a chance of our model can be under fit.



4.3 Univariate analysis

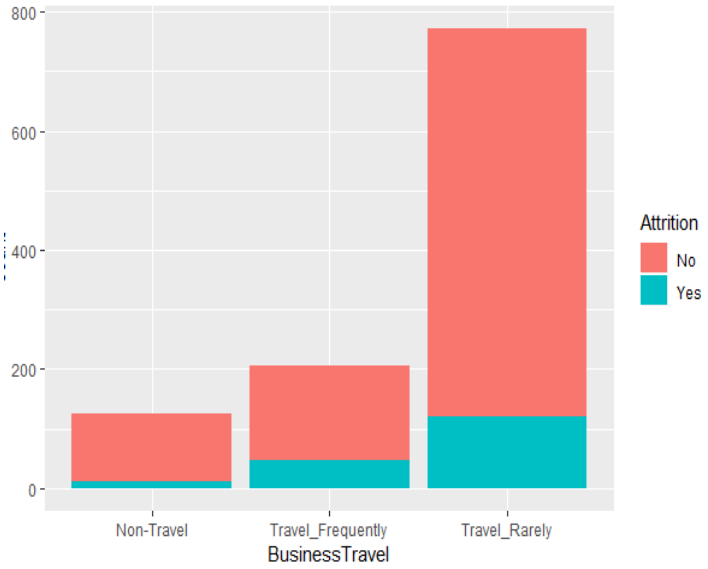
4.3.1 Age

Grouped	No of Employees
10-20	17
20-30	309
30-40	622
40-50	349
50-60	168
60 and above	5



We see that majority of employees leaving the organization are around 30 years.

4.3.2 Business Travel

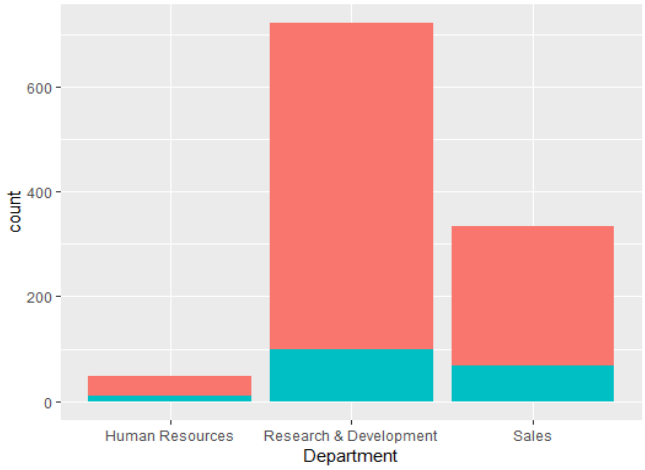


Among all the people, people who travel rarely are leaving organization most.

Business Travel	**Values**
Travel Rarely	1043
Travel frequently	277
Non-Travel	150

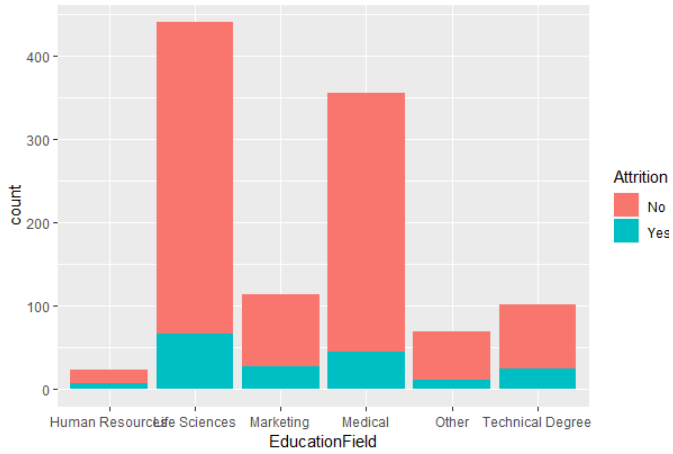
Form this, we can infer that employees who travel frequently will leave company when compared to Non-Travelers.

4.3.3 Department



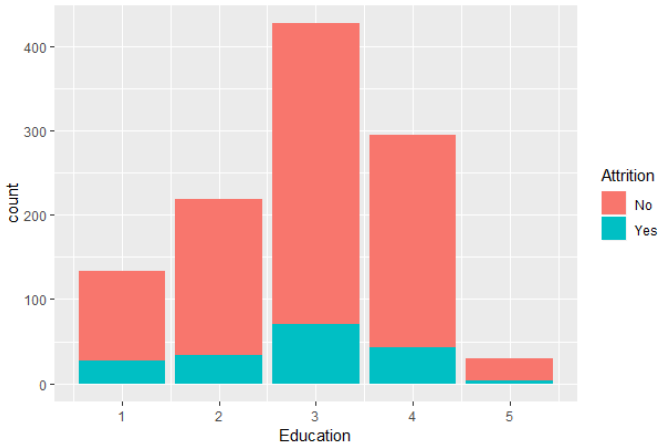
Among all the people whose attrition is “yes” is less in Human resources department because of less proportion of Human Resources in the organization.

4.3.4 Education Field



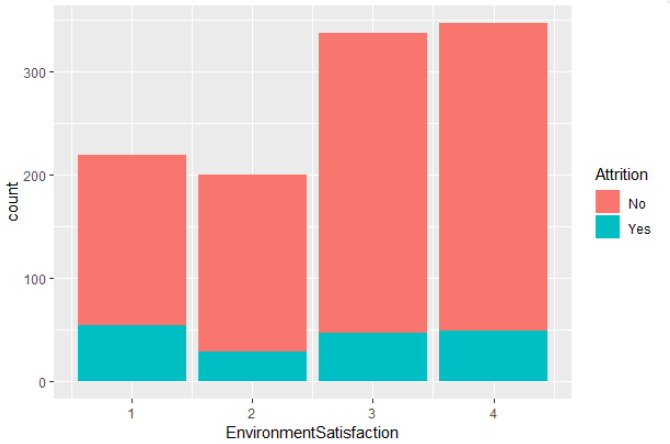
On lines of the trend in Departments, a minority of HR educated employees leave and it is majorly because of low number of people

4.3.5 Education



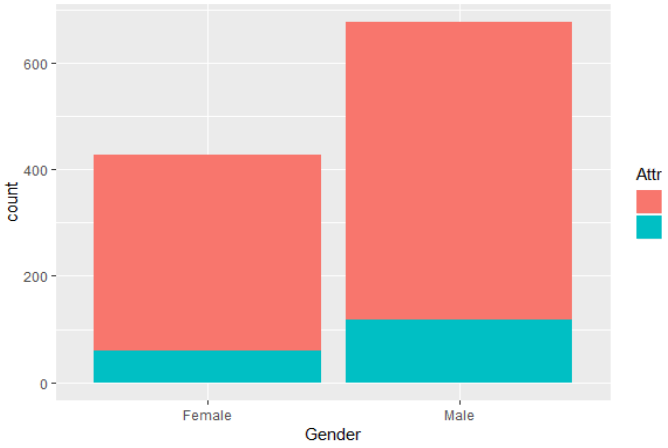
From the metadata we know that 1 ‘Below College’ 2 ‘College’ 3 ‘Bachelor’ 4 ‘Master’ 5 ‘Doctor’. Looking at the plot we see that very few Doctors attrite. May be because of less number.

4.3.6 Environment Satisfaction



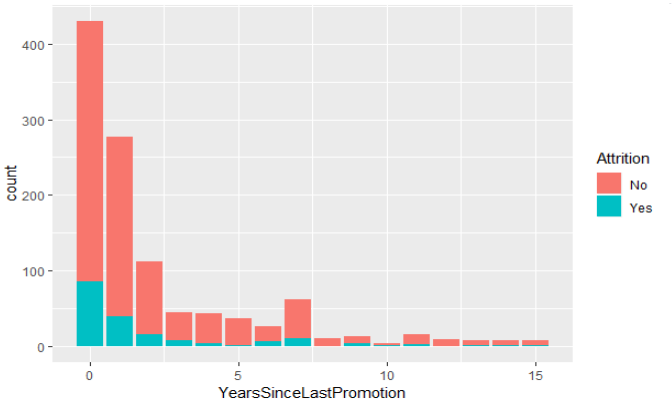
Ratings stand for - 1 ‘Low’ 2 ‘Medium’ 3 ‘High’ 4 ‘Very High’. We don’t see any distinguishable feature.

4.3.7 Gender



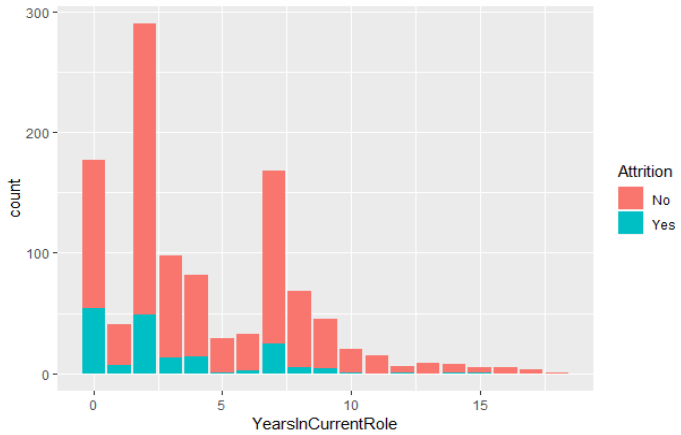
We see that majority of separated employees are Male and the reason might be because around 61% of employees in our dataset are Male.

4.3.8 Years since Last Promotion



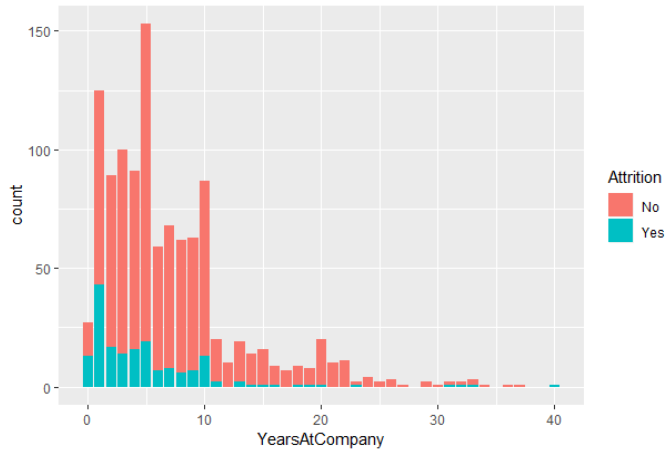
Larger proportion of people who have been promote recently have quit the organization.

4.3.9 Years since Current Role



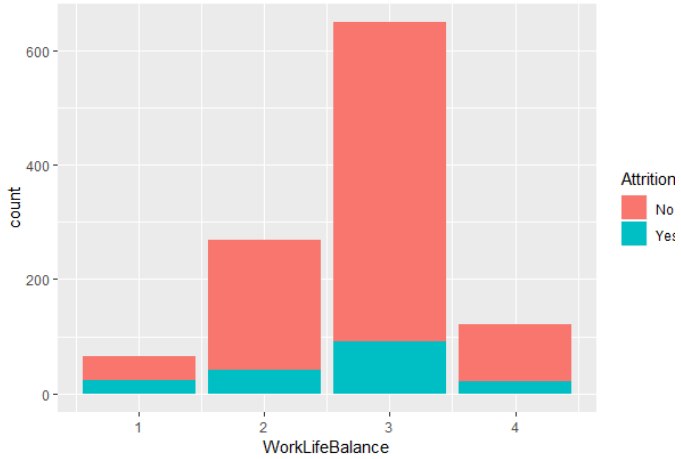
Plot shows a larger proportion with just 0 years quitting. May be a role change is a trigger for Quitting.

4.3.10 Years at Company



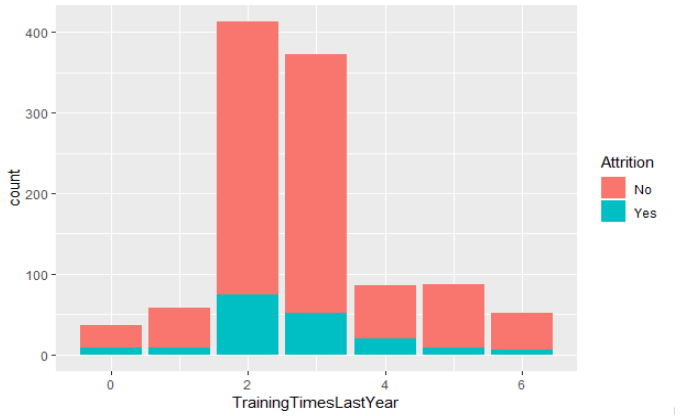
Larger proportion of new comers are quitting the organization. Which sidelines the recruitment efforts of the organization.

4.3.11 Work Life Balance

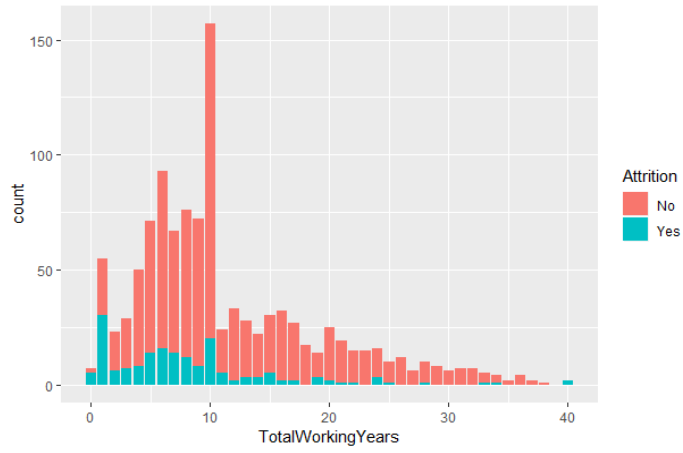


Ratings as per Metadata is 1 'Bad' 2 'Good' 3 'Better' 4 'Best'. As expected larger proportion of 1 rating quit, but absolute number wise 2 & 3 are on higher side.

4.3.12 Training Times Last Year

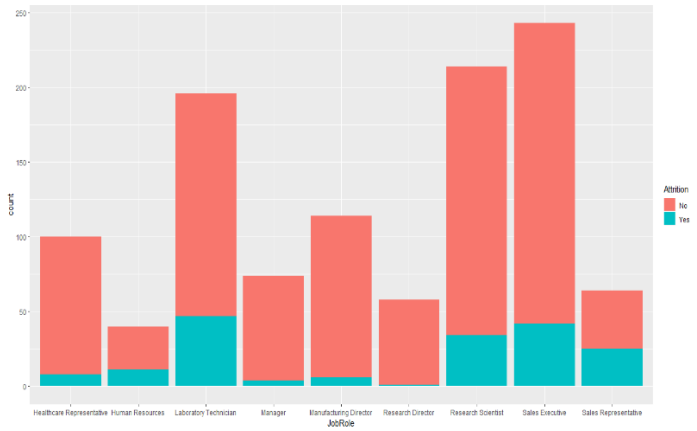


4.3.13 Total Working Years



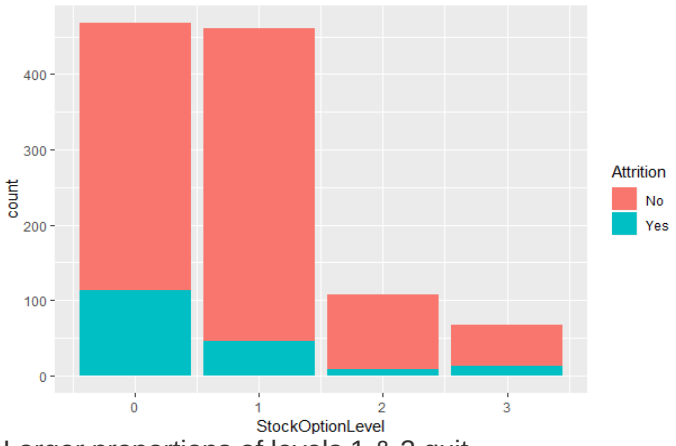
We see larger proportions of people with 1 year of experiences quitting the organization also in bracket of 1-10 Years.

4.3.14 Job Role



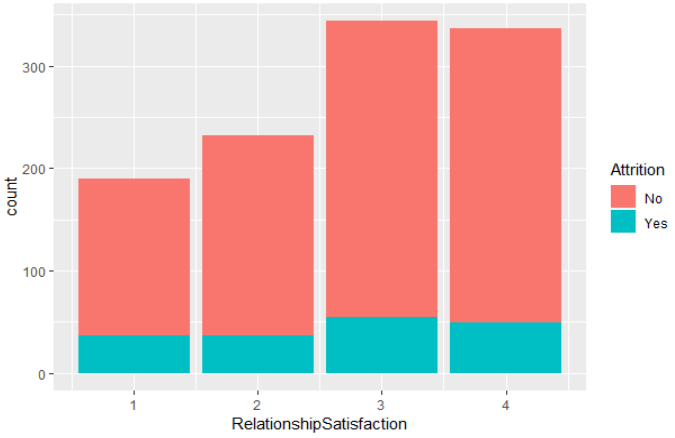
Attrition	No	Yes
JobRole		
Healthcare Representative	122	9
Human Resources	40	12
Laboratory Technician	197	62
Manager	97	5
Manufacturing Director	135	10
Research Director	78	2
Research Scientist	245	47
Sales Executive	269	57
Sales Representative	50	33

4.3.15 Stock Option Level



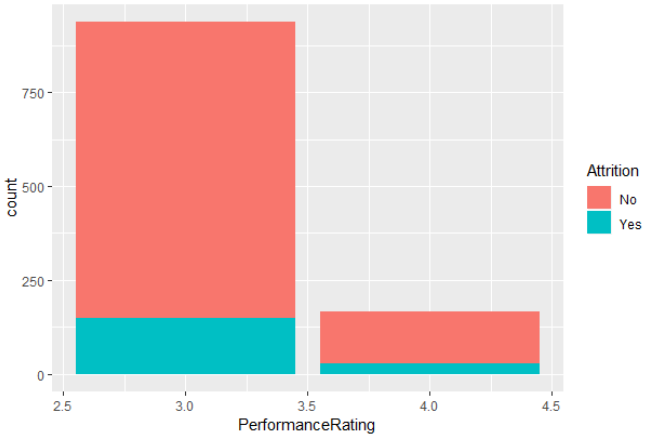
Larger proportions of levels 1 & 2 quit

4.3.16 Relationship Satisfaction



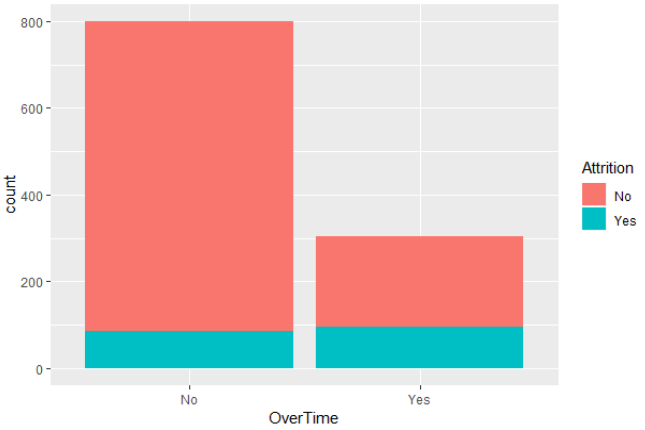
1 'Low' 2 'Medium' 3 'High' 4 'Very High'. Higher number of people with 3 or more rating are quitting. But larger proportions of 1 & 2 rating are quitting.

4.3.17 Performance Rating



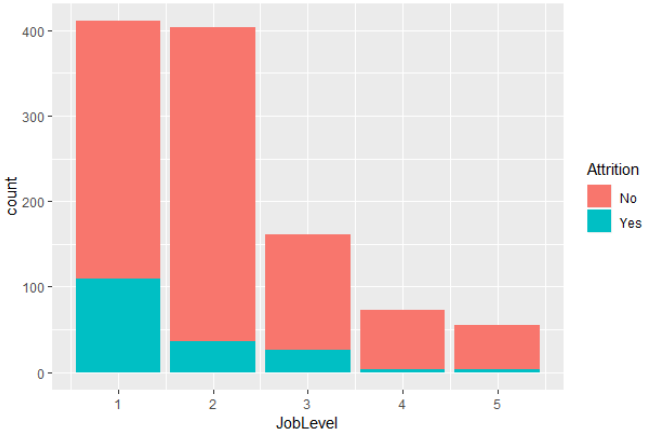
1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'. We see that we have employees of only 3 and 4 ratings. Lesser proportion of 4 raters quit.

4.3.18 Over Time



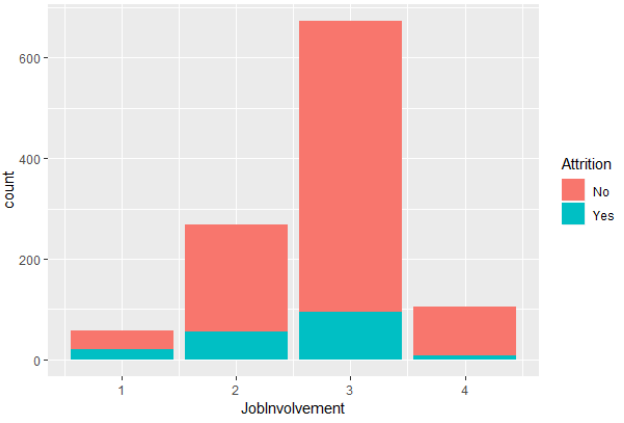
Larger Proportion of Overtime Employees are quitting.
Larger Proportion of Overtime Employees are quitting.
More than 25 % of Employees who work overtime leave the company.

4.3.19 Job Level

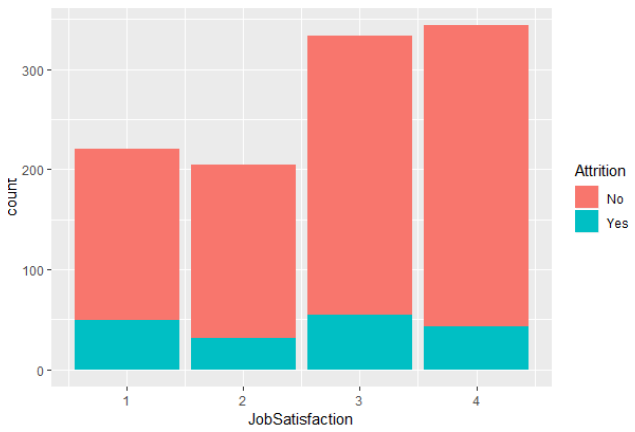


We have no metadata with regard to the numbers in Job Level. But by looking at proportion of people seems like 1 stands for entry level and 5 stands for highest level in our Dataset. By looking at plot we see that as the Job Level increases the number of people quitting decreases.

4.3.20 Job Involvement



4.3.21 Job Satisfaction

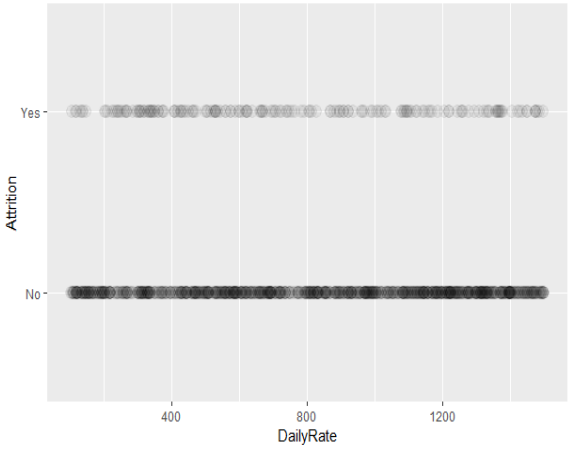


As per metadata 1 'Low' 2 'Medium' 3 'High' 4 'Very High'. We see higher attrition levels in among lower Job Satisfaction levels.

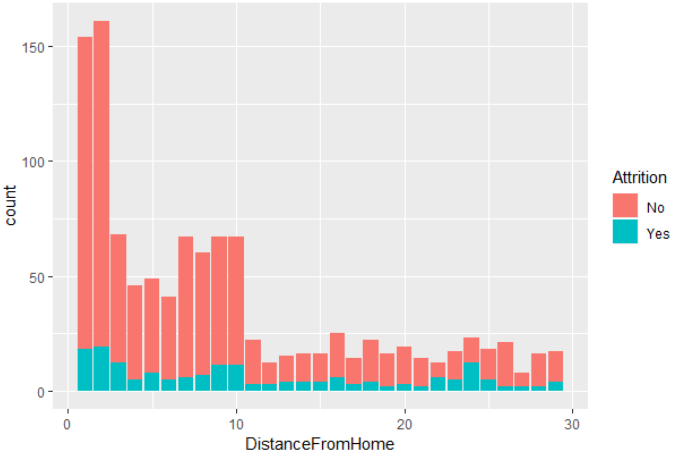
4.3.22 Hourly Rate

We don't get much inference from this. There also seems to be no straightforward relation with the Daily Rate of the employees.

4.3.23 Daily Rate

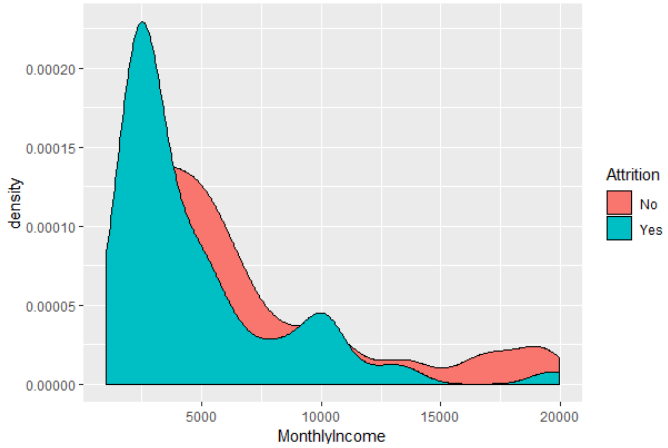


4.3.24 Distance from Home



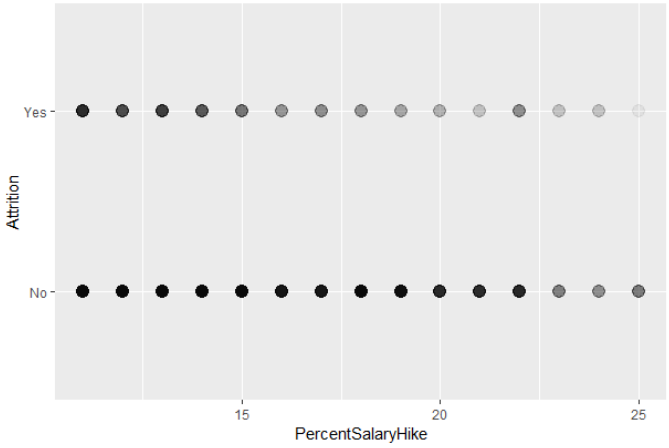
More employees leave the company if distance from home is greater than 12 kms.

4.3.25 Monthly Income



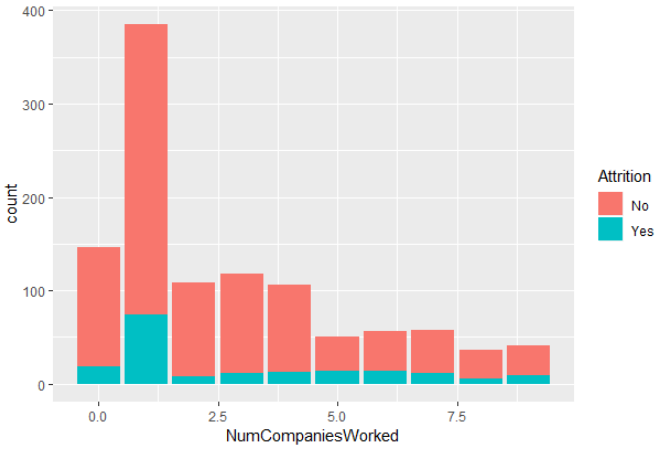
We see higher levels of attrition among the lower segment of monthly income. If looked at in isolation, might be due to dissatisfaction of income for the effort out.

4.3.26 Percent Salary Hike



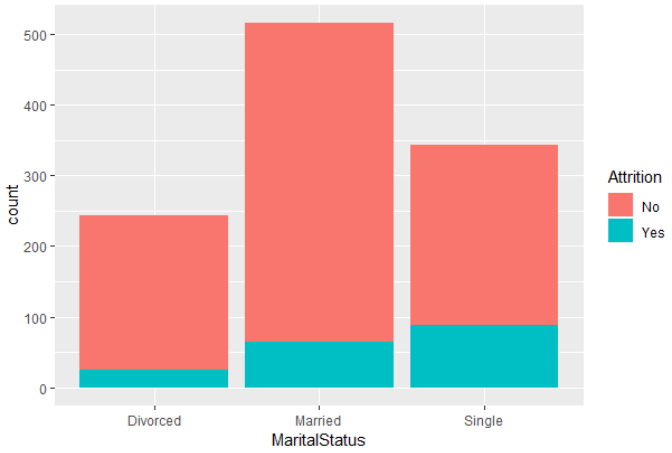
We see that people with less than 15% hike have more Chances to leave.

4.3.27 Number of companies Worked



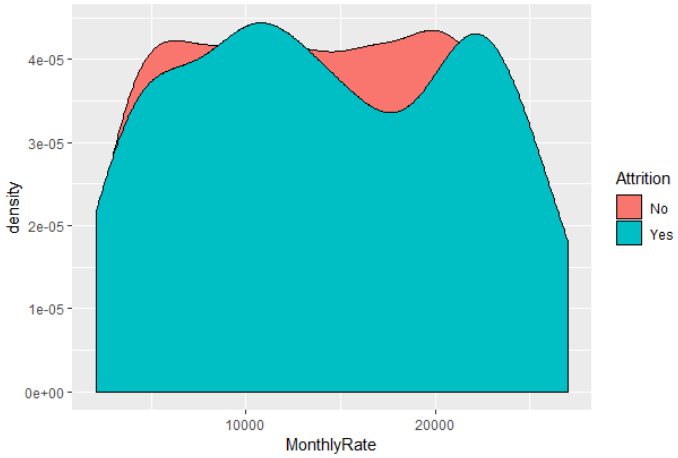
We see a clear indication that many people who have worked only in one company before quit a lot.

4.3.28 Marital status



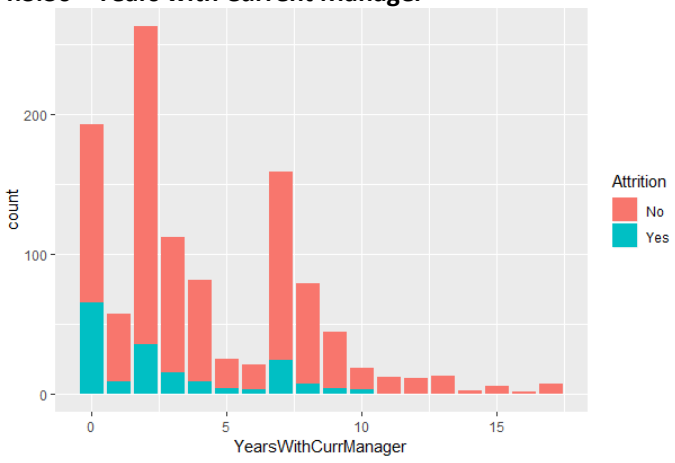
Attrition is on higher side for Single and lowest for Divorced employees.

4.3.29 Monthly Rate



We don't see any inferable trend from this. Also no straightforward relation with Monthly Income.

4.3.30 Years with Current Manager



As expected a new Manager is a big cause for quitting.

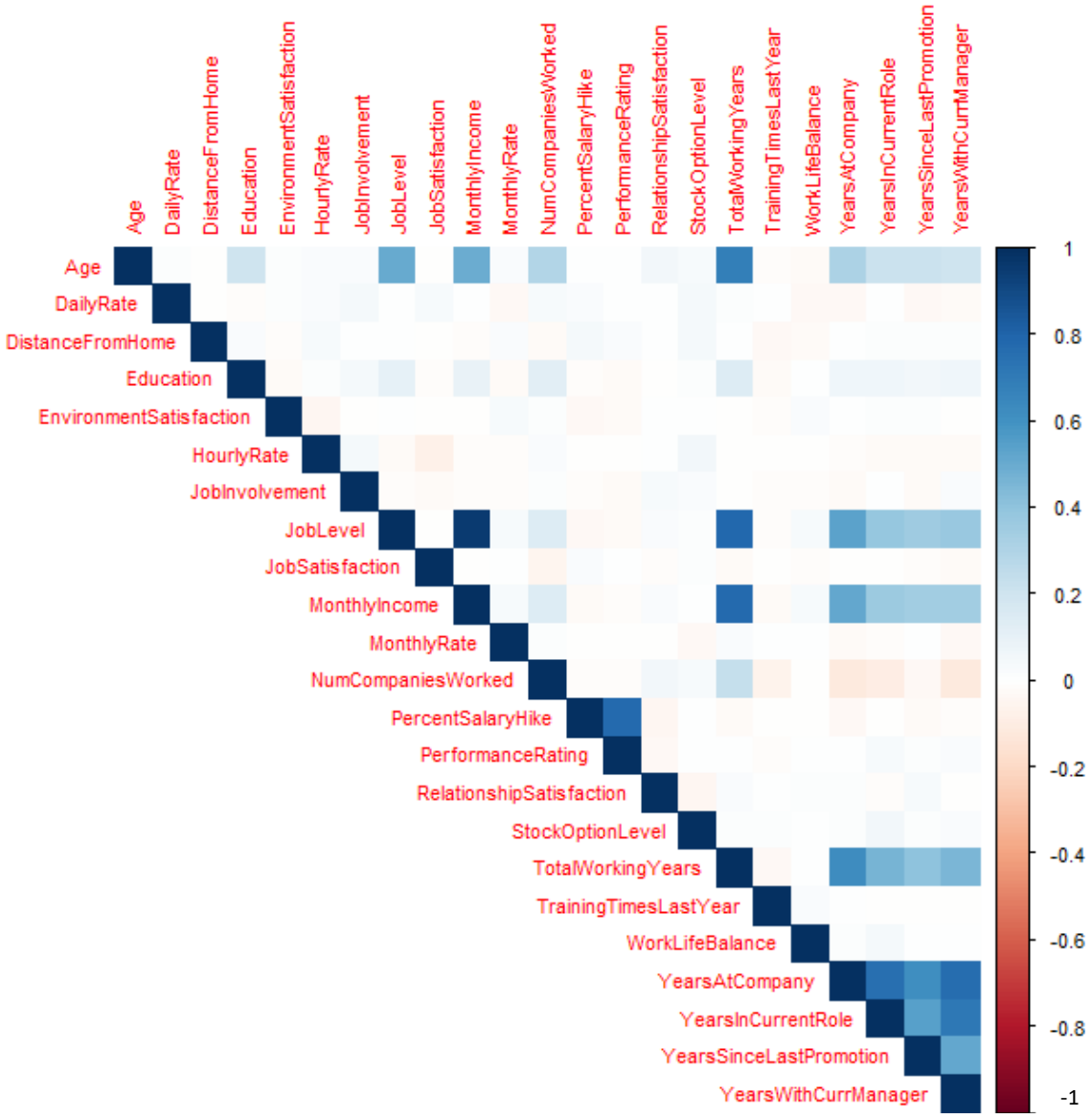
5. Inferences From Univariate Analysis

Inference from univariate Analysis:-

- For all the Rows, Employee count is equal to 1 Standard hours is equal to 80, Over18 is Y which means all are above 18. Removing 4 variables including Employee Number.
- Many of the Employees Education is 3 and JobLevel is 1 and 2. Monthly income of many of the Employees is below 2500.
- For many of Employees, this is the second company (Number of Companies Worked is more for 1) and total working years is less than 10 years for most of the Employees and years at this company is less than 5 years.
- Most of the Employees Performance Rating is 3 and many of Employees' Percent Salary Hike is less than 15%.

So, Education Field, Gender, Department, Trainingtimessincelastyear, performance rating and Education Field are not strong predictors and we will not be including these variables

6. Checking if there is Multi-Co linearity - High Correlation between



- We can see that Age and *TotalWorkingYears* are highly correlated (we know that total working years is dependent on age).
- We can see that Monthly Income and *TotalWorkingYears* are highly correlated.
- As Age Increases, *TotalWorkingYears* increases and As Total working year's keeps increasing, Monthly income will increase. So, we will not be considering these variables.
- *YearsAtCompany* is correlated with *YearsInCurrentRole*, *YearsWithCurrManager* and *TotalWorkingYears*. So, we will not be considering *YearsAtCompany*. We will not be considering *YearsWithCurrManger* as there is correlation with *YearsIncurrentRole*.

7. Modelling with Decision Tree

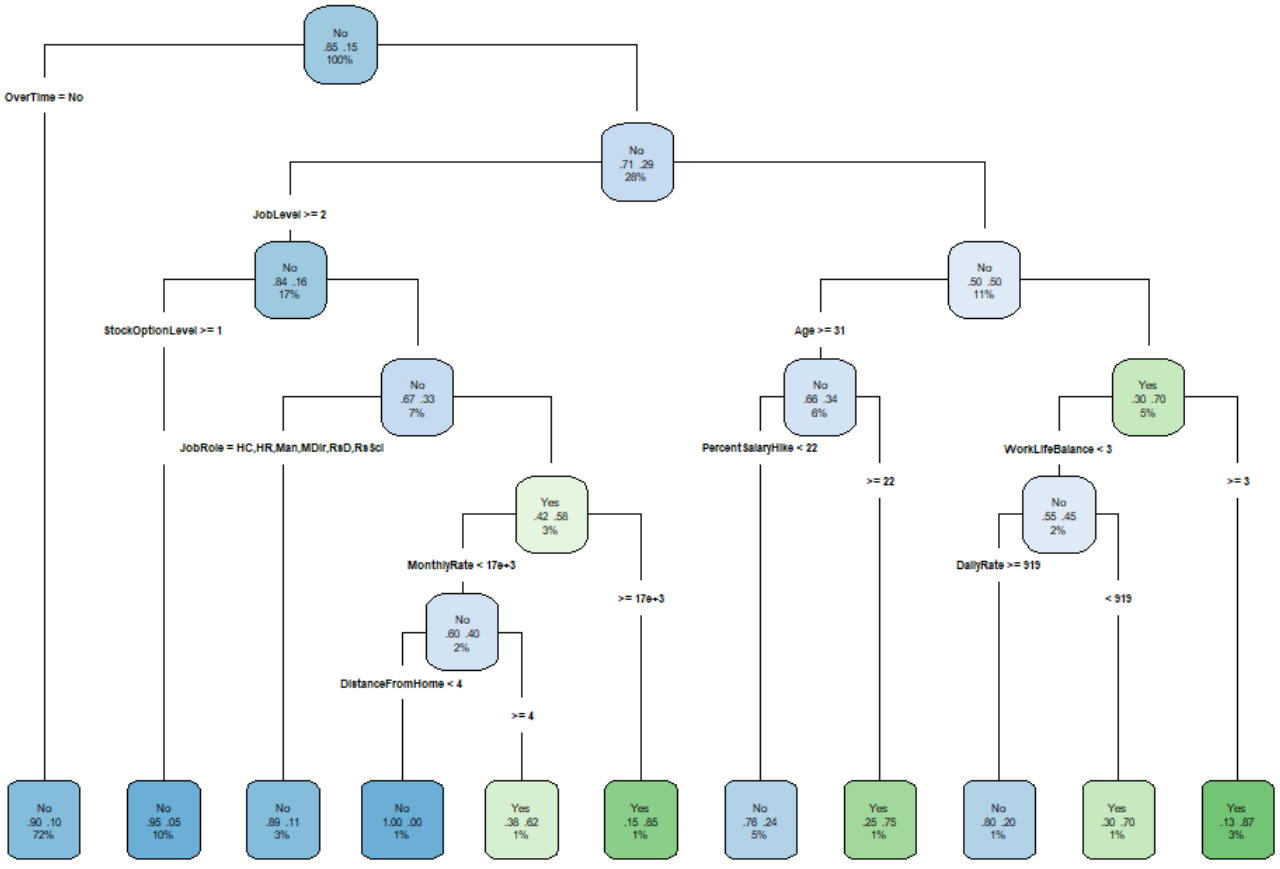
Area under the curve: 0.5961

7.1 Confusion matrix

Actual \ Predicted	NO	Yes
NO	0.96143251	0.03856749
YES	0.76923077	0.23076923

A rather poor AUC with a rather poor sensitivity. It seems that building a single tree will not bring us anywhere. However, while not being useful in general, such a model can nevertheless help us see some patterns. Let us plot the tree and see if we can find any.

7.2 Decision Tree



7.3 Insights of Decision Tree

The top 5 factors that influence the attrition seem to be:

- Overtime,
- Monthly income,
- Job level,
- Age,
- Number of companies worked for

Two of these are already familiar to us from our EDA and decision tree plot - it seems that we should indeed do something about those who work overtime and then leave and those who have a low monthly income (which is probably also linked to the job level).

We should also delve a bit more into the matter of age & number of companies the person worked for. Isn't that simply linked to people who retire, and who e.g. probably worked for many companies throughout their life? Or to the fact that we frequently hire freelancers for some temporary positions? If not, what could be wrong there? What policies and/or services are we lacking?

Last but not least, the fact that all three variables linked (directly or indirectly) to work-life balance (distance from home, business travel, and work-life balance as such) have their place among the top 20 variables could also be a sign that something should be done in this area. Remember, we've already observed this pattern during the visualization phase.

8. Suggested Action

- The main general reason behind attrition is most likely the effort-reward imbalance. In this case, this mostly applies to people who are working overtime and who in many cases have a relatively low salary - it should be checked whether there is an effective overtime policy in our company;
- Our simple decision tree shows that further solutions may not lie uniquely in people getting higher salaries (or their overtime pay). Those with relatively higher salaries may be interested in something more than just a paycheck, and might still leave if they do not feel part of the company (e.g. if they don't have any stock options, or if they don't have access to trainings);
- We have also found that different facets of work-life balance might represent an issue for our employees (a finding supported by visualizations and (at least to some extent) our best algorithm). One of the things that should be checked is e.g. whether there is a lack of certain teleworking arrangements in our company;
- There seems to be a link between attrition and age as well as the number of companies worked for. At this point, we cannot provide more information and it would be necessary to delve deeper into our dataset, e.g. to ascertain whether this is not simply linked to retirements or to see whether there is an unfair treatment of certain age groups and whether specific part of our workforce is in need of an intervention (e.g. more job security, upskilling, etc.).
- If we take our "test" set as an example of IBM's current workforce, we can say that the job role with highest probability of attrition is sales representative - something should be definitely done about that, and we could explore further what exactly.
- Last but not least, if we would be given a new dataset of our employees, we could calculate probabilities and see which employees exactly are prone to leaving - with an algorithm that outperformed standard algorithms (e.g. Random Forest)!