Factors Impacting Employee Satisfaction

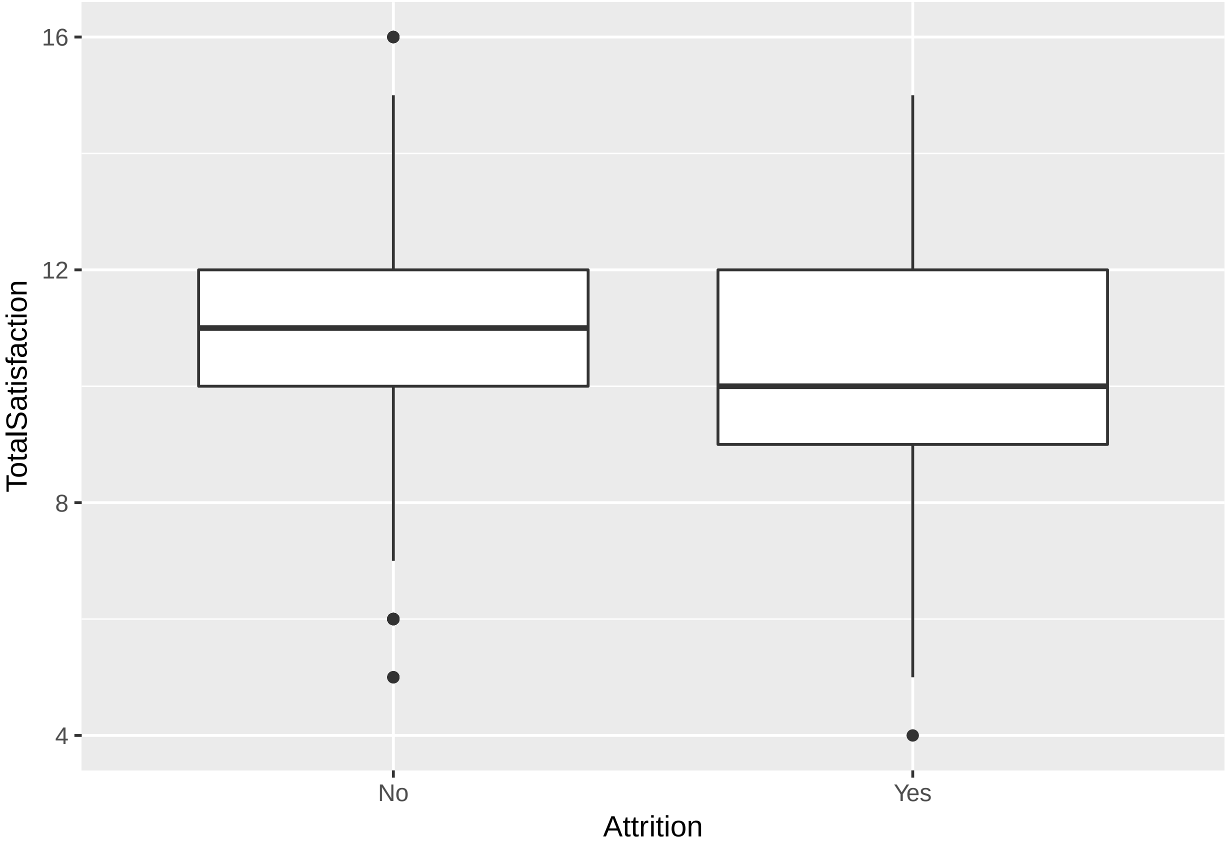
Hai Wen Chen, Avi Mago, William Marshall, Haochen Song, Siddharth Srinivasan Swamy, Xuying Zhong

September 25, 2017

Contents

Introduction

test referencing Medina (2012) ## The relationship between Attrition and Total Satisfaction



Dataset Description and Descriptive Statistics

This dataset described the employees of IBM, providing a total of 1470 observations (employees) and 35 variables (information about the employee). In our analysis, the colomns EmployeeCount and EmployeeNumber are not useful, and the value in Over18 and StandardHour are the same with all employees (Yes and 80 respectively), so we got rid of thses four colomns to make our analysis more efficient. Meanwhile, we added a new colomn called "TotalSatisfaction", which is the sum of EnvironmentSatisfaction,JobSatisfaction,RelationshipSatisfaction and WorkLifeBalance. In addition, the Education colomn uses integer between 1 to 5 to refer to the level of education of this employee, they mean below college, college, bachelor, master, doctor respectively. To catogrize the education level, we change the integer incicating the level in to factor. We did the same thing to JobLevel, StockOptionLevel and TrainingTimeLastYear. Then the variables can be divided into three groups: Personal Information, Job Information, Satisfaction to the Job in Total.

Table 1: Personal Information

Variable	Type	Description
Age	integer	The Employee's age
DistanceFromHome	integer	The distance from home to work
Education	factor	Level of education (1 'Below College', 2 'College', 3 'Bachelor', 4 'Master', 5 'Doctor')
EducationField	factor	The subject of the employee's education (Human Resources, Life Sciences, Marketing, M

Variable	Type	Description
Gender	factor	Gender of this employee
MaritalStatus	factor	Marital status of this employee (Divorced, Married, Single)
NumCompaniesWorked	integer	The number of companies this employee has worked
TotalWorkingYears	integer	Total number of years this employee has worked since graduation

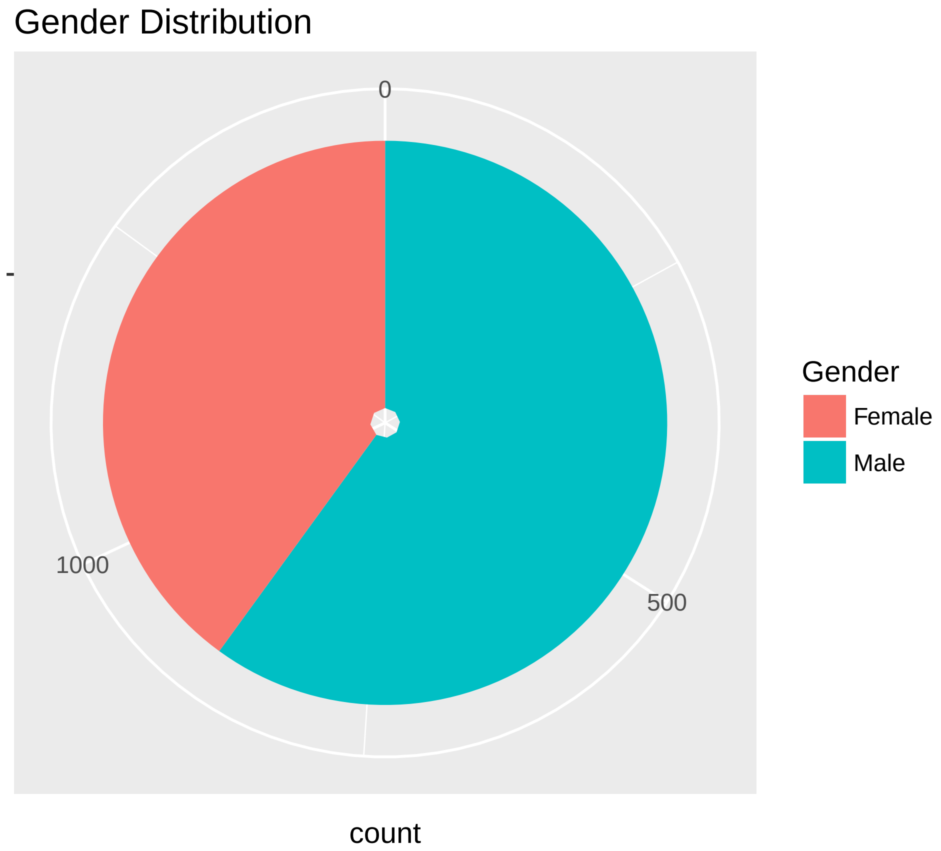
Table 2: Job Information

Variable	Type	Description
Attrition	factor	If this employee left the company
BusinessTravel	factor	The frquency of business travel
DailyRate	integer	sth
Department	factor	The department of this employee (Human Resource, Research & Development and Sal
HourlyRate	integer	sth
JobInvolvement	factor	1 ‘Low’, 2 ‘Medium’, 3 ‘High’, 4 ‘Very High’
JobLevel	factor	The level of this employee’s job
JobRole	factor	The position of this employee (Sales Executive, Research Scientist, Laboratory Techni
MonthlyIncome	integer	The salary of this employee
MonthlyRate	integer	sth
OverTime	factor	If this employee works over time (Yes, No)
PercentSalaryHike	integer	The percentage of salary hike
PerformanceRating	integer	1 ‘Low’, 2 ‘Good’, 3 ‘Excellent’, 4 ‘Outstanding’
StockOptionLevel	factor	The amount of stock this employee process
TrainingTimesLastYear	factor	The length of training the employee took last year
YearsAtCompany	integer	The number of years this employee has been in the company
YearsInCurrentRole	integer	The number of years this employee has been in this position
YearsSinceLastPromotion	integer	The number of years since last promotion
YearsWithCurrManager	integer	The number of years this employee has been with current manager

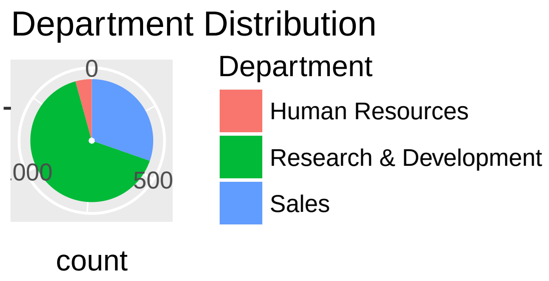
Table 3: Satisfaction to the Job in Total

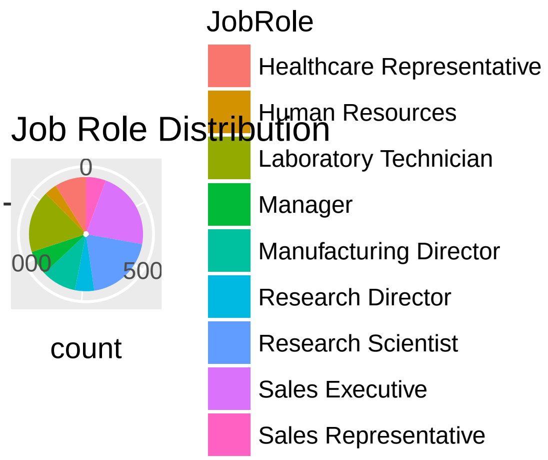
Variable	Type	Description
EnvironmentSatisfaction	integer	Satisfaction to the environment (1 ‘Low’, 2 ‘Medium’, 3 ‘High’, 4 ‘Very High’)
JobSatisfaction	integer	Satisfaction to the job (1 ‘Low’, 2 ‘Medium’, 3 ‘High’, 4 ‘Very High’)
RelationshipSatisfaction	integer	Satisfaction to the relationship (1 ‘Low’, 2 ‘Medium’, 3 ‘High’, 4 ‘Very High’)
WorkLifeBalance	integer	The work life balance rate (1 ‘Bad’, 2 ‘Good’, 3 ‘Better’, 4 ‘Best’)
TotalSatisfaction	integer	The sum of EnvironmentSatisfaction,JobSatisfaction,RelationshipSatisfaction and Work

This dataset covered 1470 employees and according to the pie chart, male make up 2/3 of the total employees.



It includes information about employees from 3 different departments in 7 different roles. 2/3 of the recorded employees are from Research & Development department, employees from Sales took up most of the remaining part and there are only 63 employees come from Human Resources department recorded in this dataset.





Statistical Tests

```
prop.test(table(IBM[,"Gender"], IBM[,"Attrition"]), alternative = "greater")
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  table(IBM[, "Gender"], IBM[, "Attrition"])
## X-squared = 1.117, df = 1, p-value = 0.1453
## alternative hypothesis: greater
## 95 percent confidence interval:
##  -0.01113656  1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.8520408 0.8299320
```

The proportion of female staying in the company is 85.20% The proportion of male who stay in the company is 82.00%. So the proportion of male and female who stay in the company is the same according to the proportion test.

Hypothesis

Method

<<<<<<< HEAD <<<<<<< HEAD ======= # Analysis

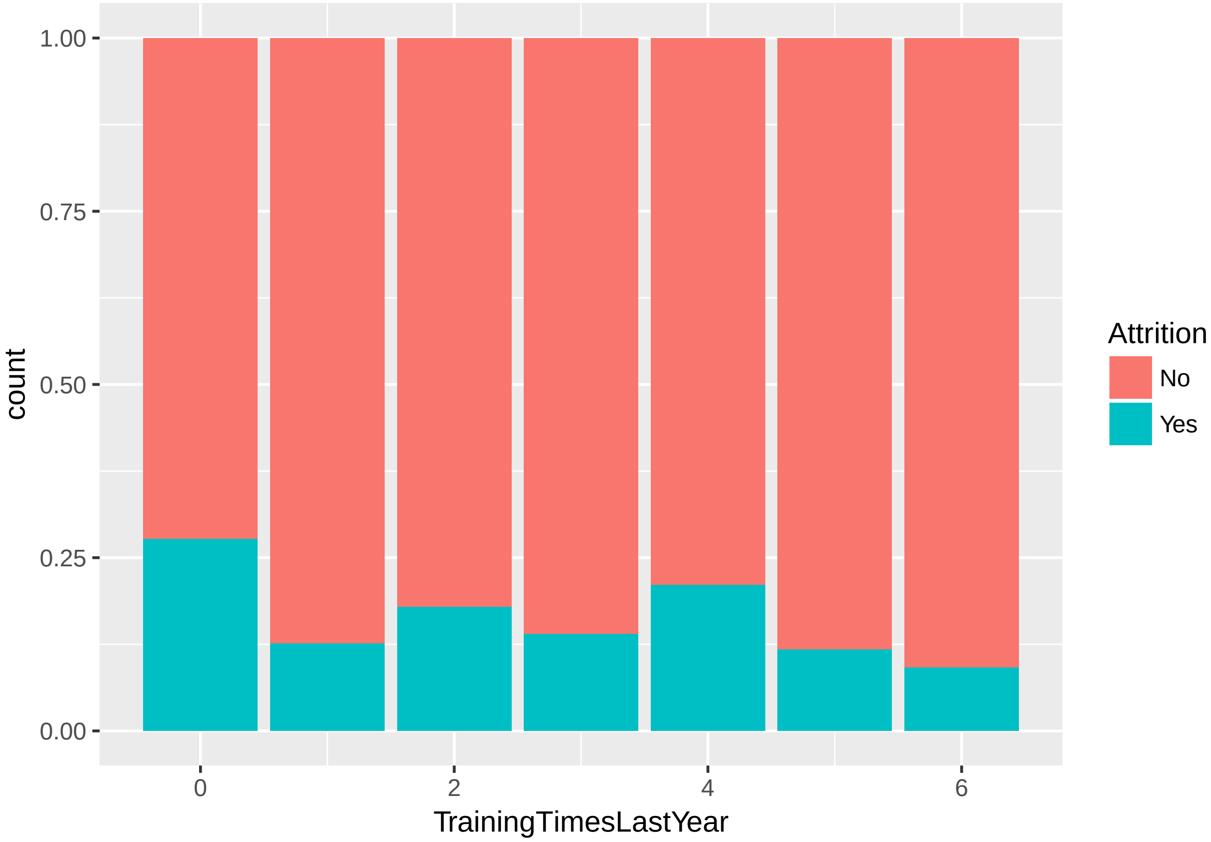
Converting variables into Factor Variables

IBM$Education < -ordered(IBM$Education,levels=c(“1”,“2”,“3”,“4”,“5”)) IBM$JobSatisfaction <
-ordered(IBM$JobSatisfaction,levels=c(“1”,“2”,“3”,“4”)) IBM$JobLevel < -ordered(IBM$JobLevel,levels=c(“1”,“2”,“3”,“4”)
IBM$JobInvolvement < -ordered(IBM$JobInvolvement,levels=c(“1”,“2”,“3”,“4”)) IBM$JobLevel <
-ordered(IBM$Joblevel,levels=c(“1”,“2”,“3”,“4”,“5”)) IBM$StockOptionLevel < -ordered(IBM$StockOptionLevel,levels=c(“
IBM$TrainingTimesLastYear < -ordered(IBM$TrainingTimesLastYear,levels=c(“0”,“1”,“2”,“3”,“4”,“5”,“6”))
IBM$RelationshipSatisfaction < -ordered(IBM$RelationshipSatisfaction,levels=c(“1”,“2”,“3”,“4”))
IBM$WorkLifeBalance < -ordered(IBM$WorkLifeBalance,levels=c(“1”,“2”,“3”,“4”))

Plots

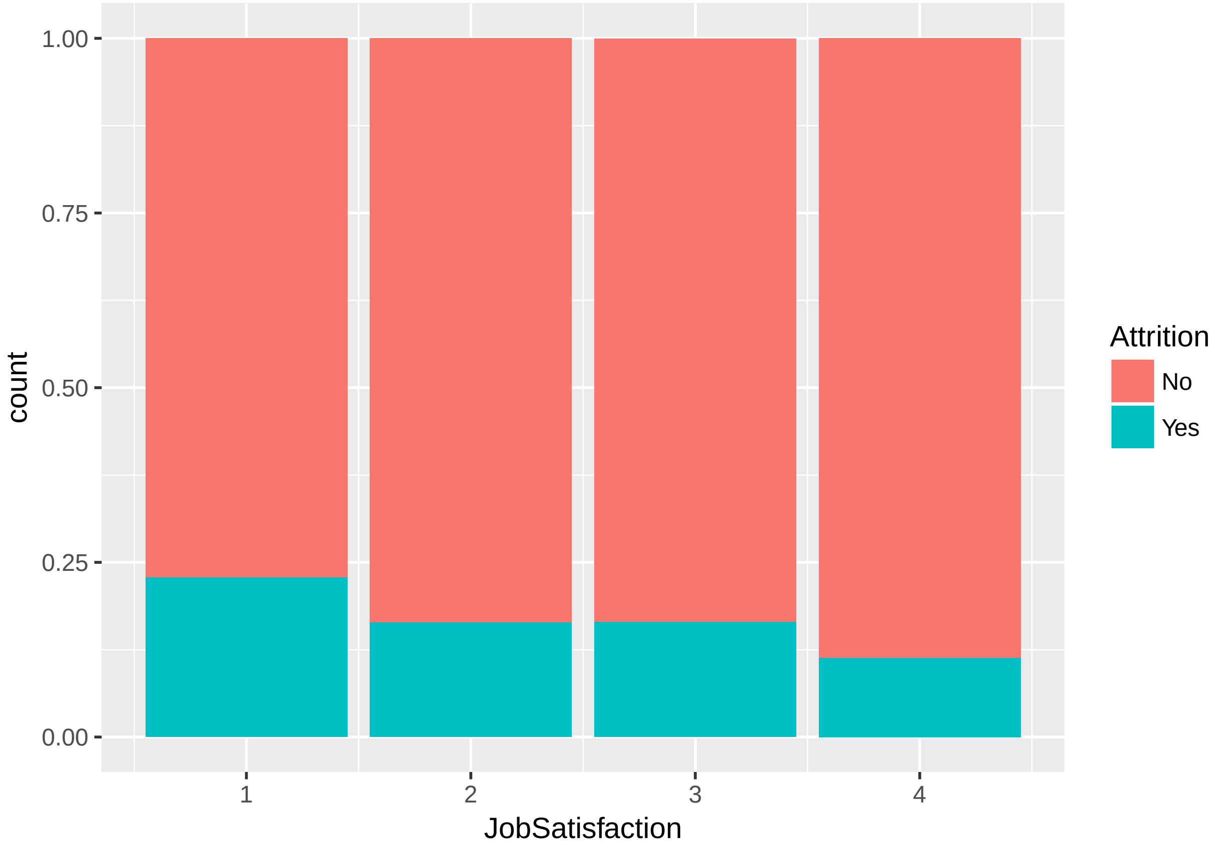
In this part, we are going to plot graphs between attrition and other different factors in order to explore their relationships and further understand which factor actually affect employees’ choice on Attrition.

Training Time Last Year (Yes vs No)



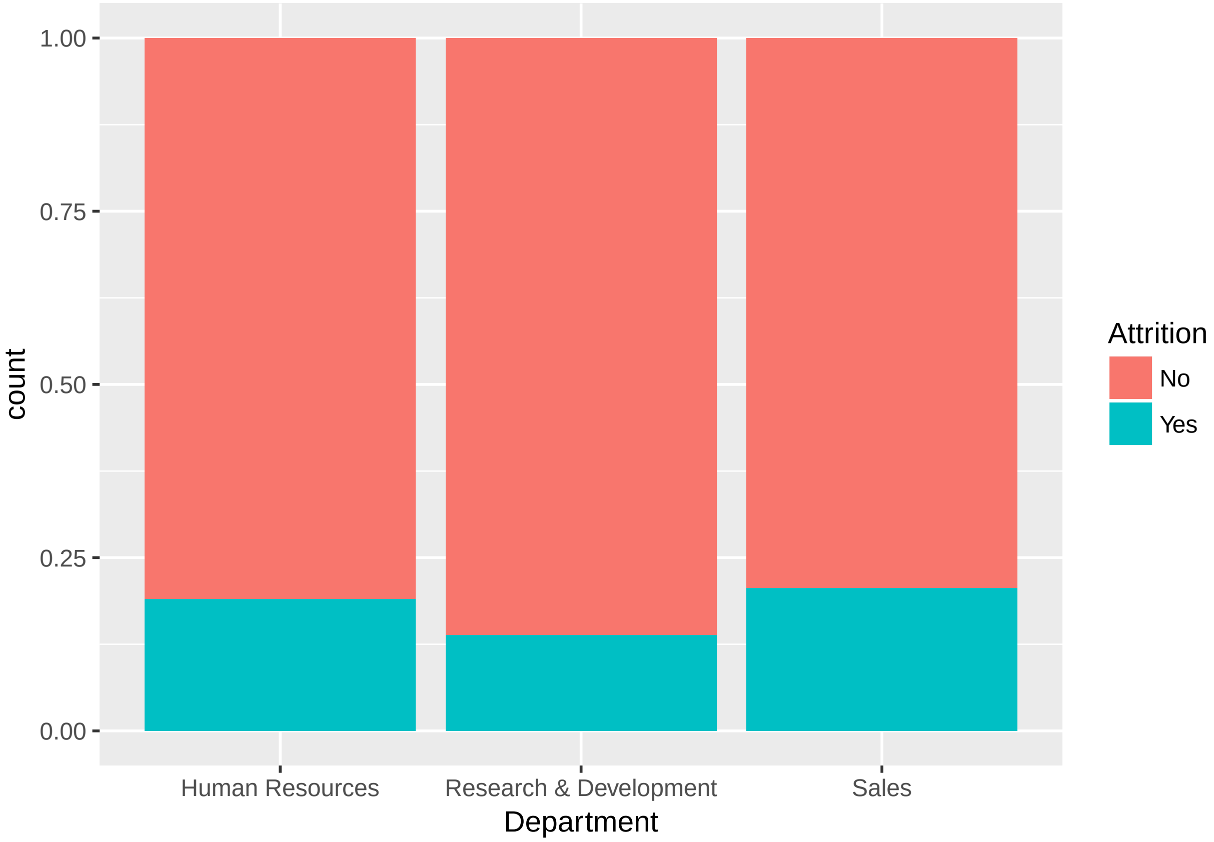
The bar chart above shows a roughly decreasing trend between the training time and the amount of attrition, the longer training time leads to fewer attrition. In addition, about 75 percent of employees keep stay at the company without being affected by this factor.

Job Satisfaction (Y vs N)



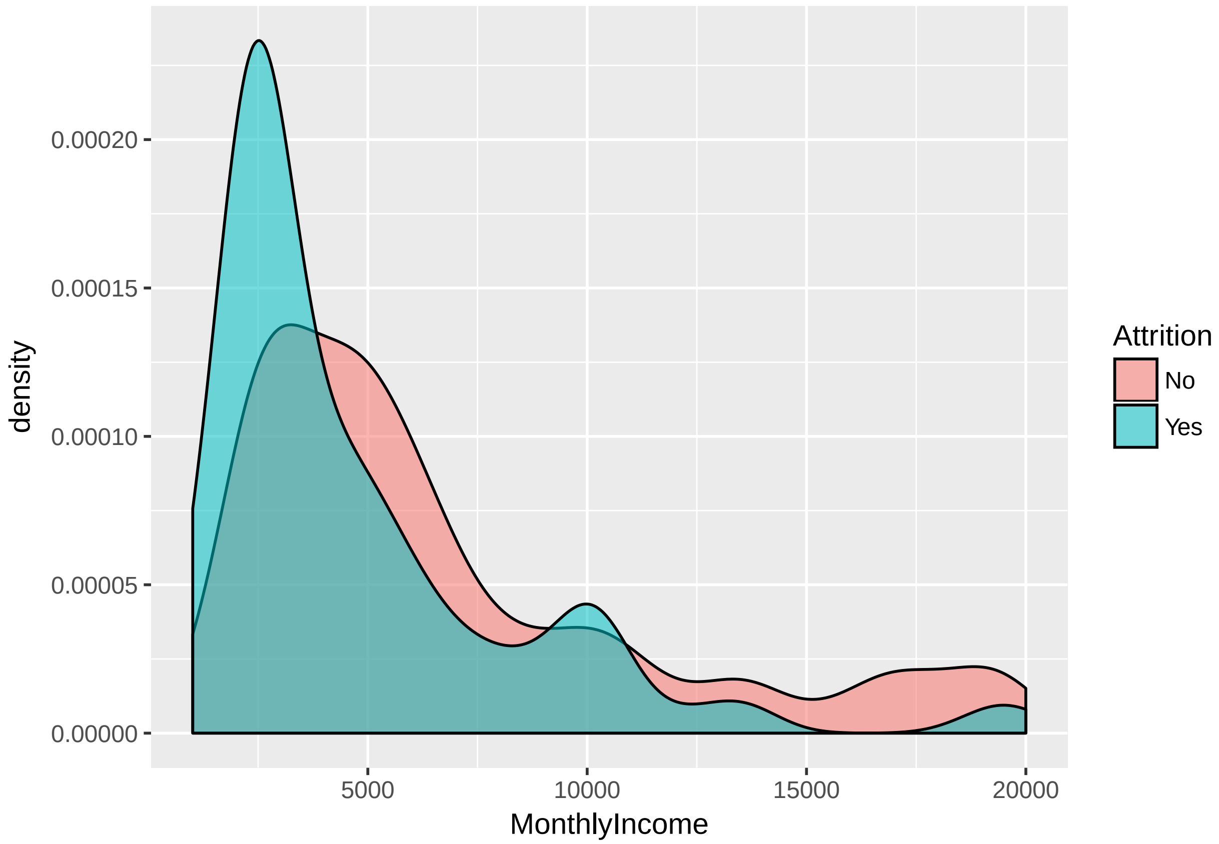
Similar to traing time, JobSatisfaction is also a factor that influece people's decision on whether to leave or not. As job satisfaction increases from 1 to 4, the percentage of attrition decreases from 24% to 10%, which indicates improving job satisfaction may help prevent employees from leaving the company.

Department wise Attrition (Y vs N)



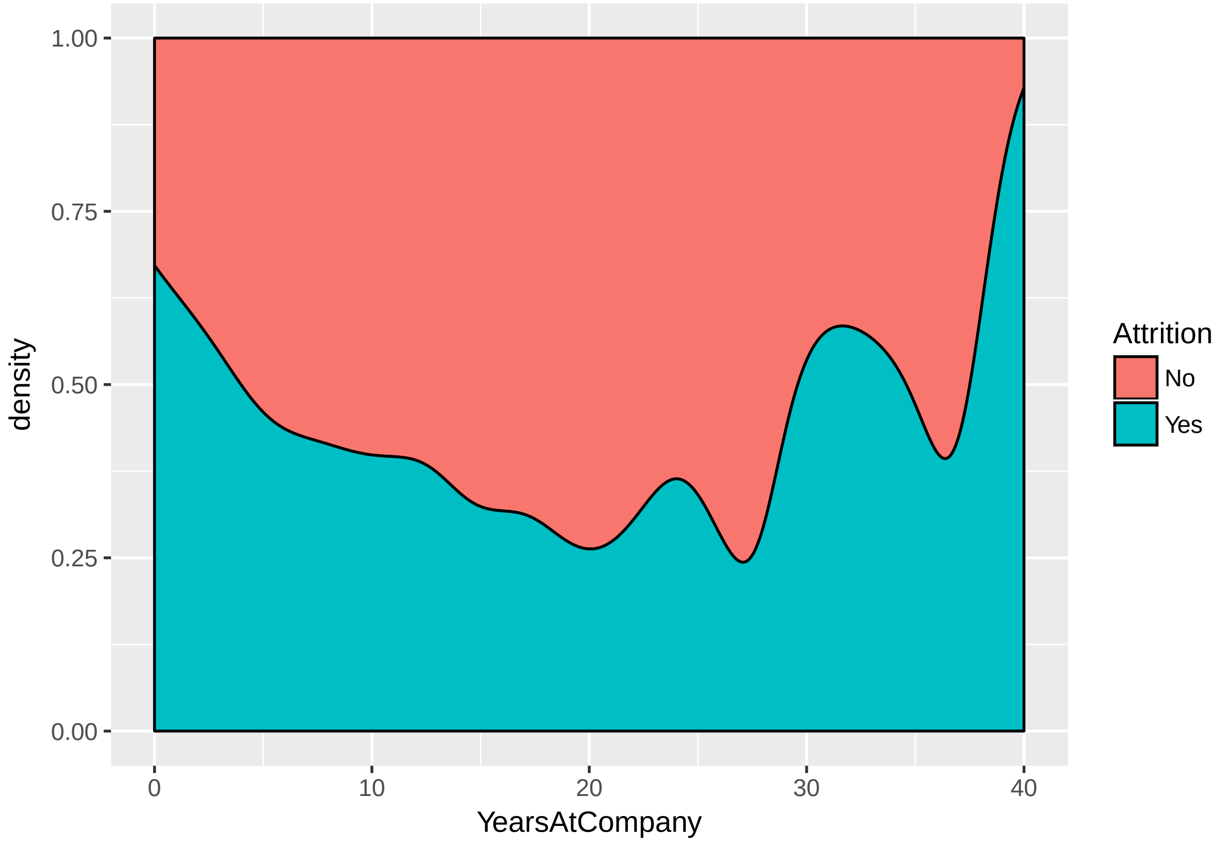
Based on the above bar chart, it can be concluded that Research & Development department is the most stable department with only around 14% of employees leaving, while the proporion in Sales and Human Resources Departments are about 10% higher than it.We can mainly focus on this two departments' analysis if the company do not have enough money for improvement.

Monthly Income Density (Y vs N)



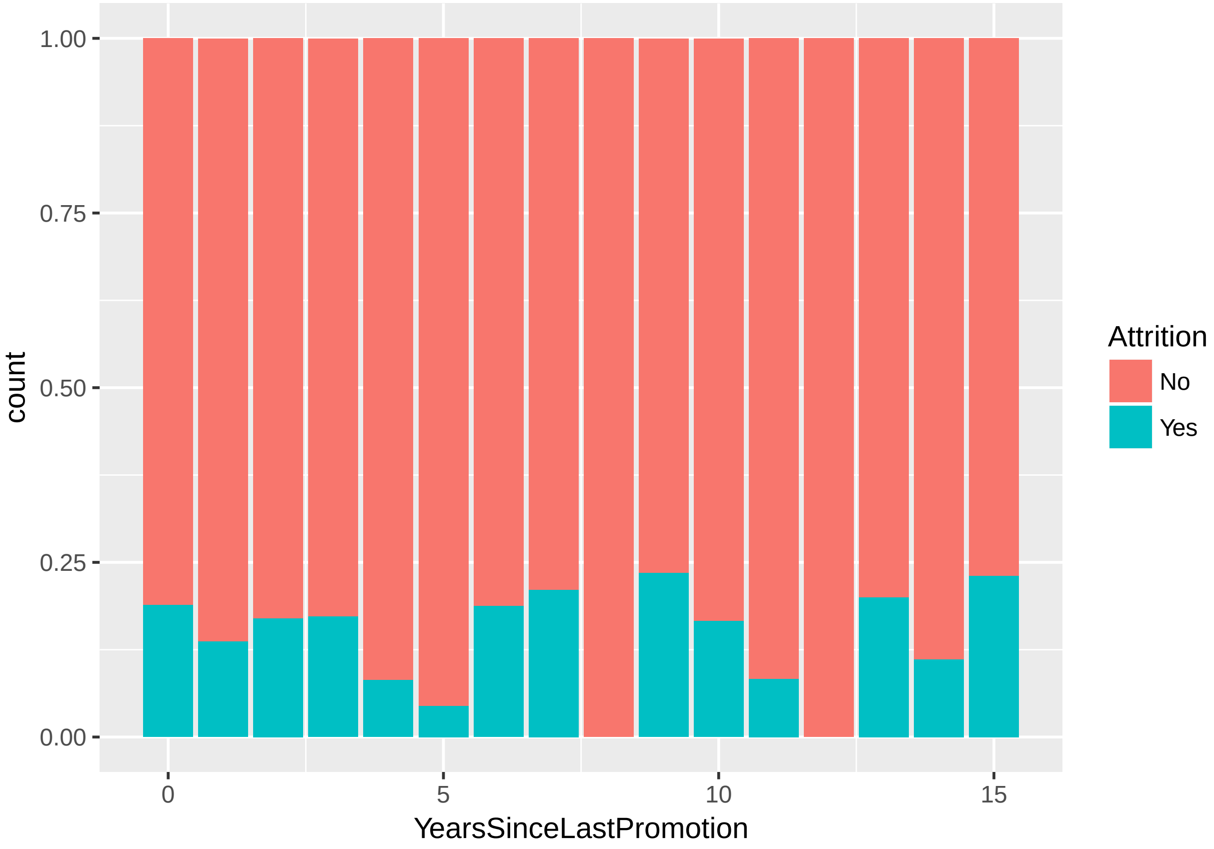
As we can see from this density plot on monthly income, it is highly right skewed, which implies most employees' incomes are aound 2500 per month. There is a sharp decrease after the peak reached at 2500, additionally, the density of employees who choose to leave is higher than who want stay when their salary are lower than 4000($).

Years At Company Density (Y vs N)



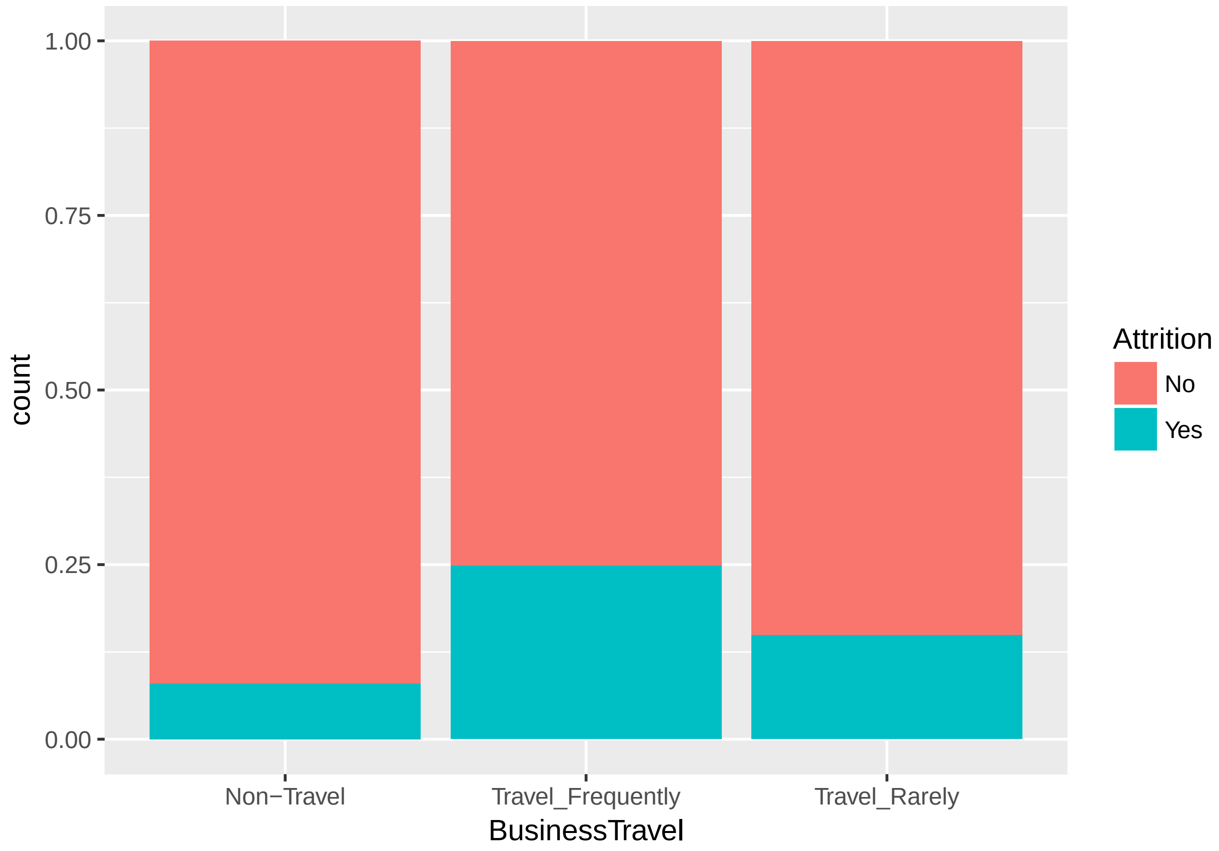
In term of years employees have stayed at company, the percentage of attrition smoothly goes down during the first 20 years.However, in the following 20 years, it then inscreases to around 90 percent with fluctuations in between. To sum up, People are more likely to leave the company during the first few years or after staying for more than 30 years.

Years Since Last Promotion (Y vs N)



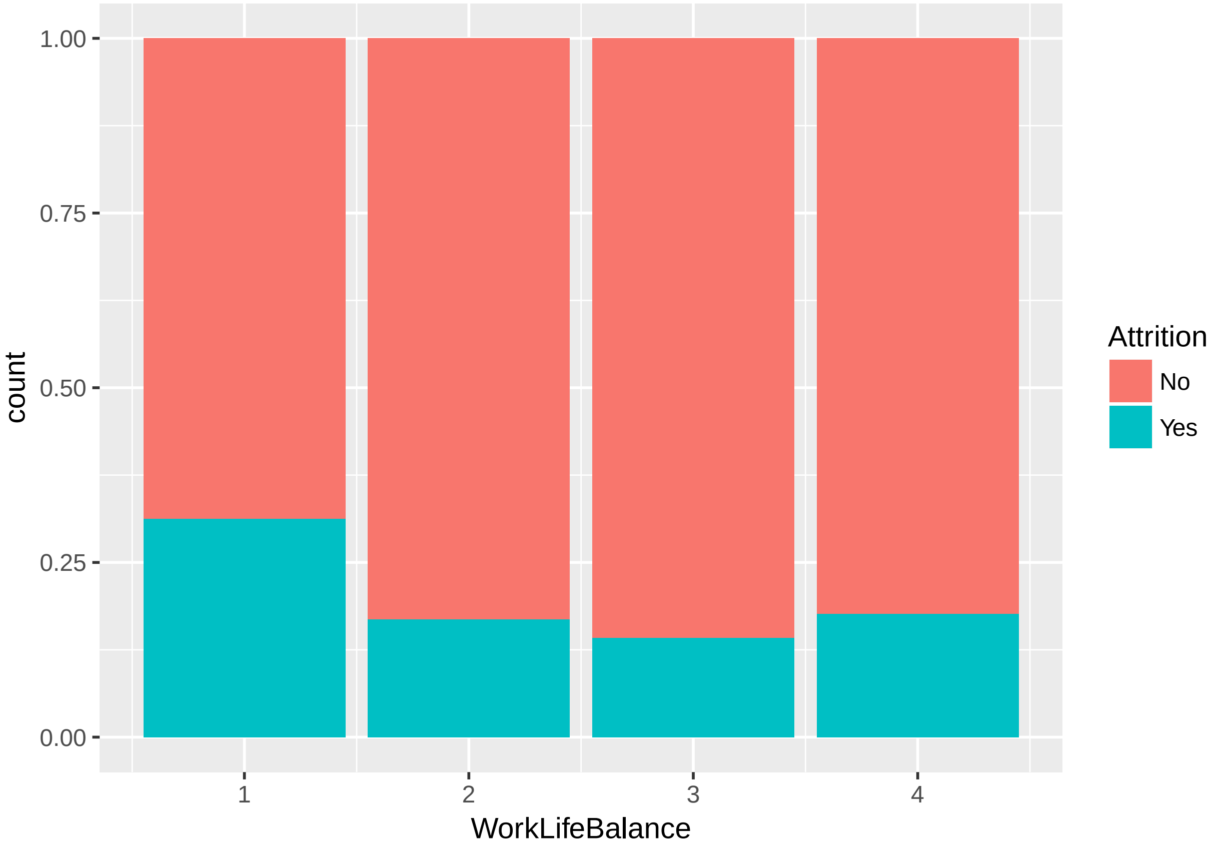
The bar chart above shows there is no significant relationship between years since last promotion and the attrition. The propotion for leaving fluctuates around 20 percent.Two special years occur at year 8 and 12 with no attrion.

Business Travel (Y vs N)



In this business travel bar chart, only 8 percent of employees want to leave company if they are in non-travel roles, but high attrition occurs at about 25 percent when they need travel frequently, which indicates that business travel may have positive impact on company's attrition.

Work Life Balance (Y vs N)



In terms of work life balance, around 30 percent of employees choose to leave the company if they have a bad balance. However with the level increases to 2 and more, the attrition percentage drops dramatically to half of level 1. Therefore, improvement of work life balance can also be a significant factor.

Colusion for Attrition plot analysis

So far we have plotted eight graphs aiming to find out the relationship between attrition and each other variables. From the analysis above, we can see JobSatisfaction, BusinessTravel and WorkLifeBalance are three most significant factors that have strong impact on employees attrition choices.

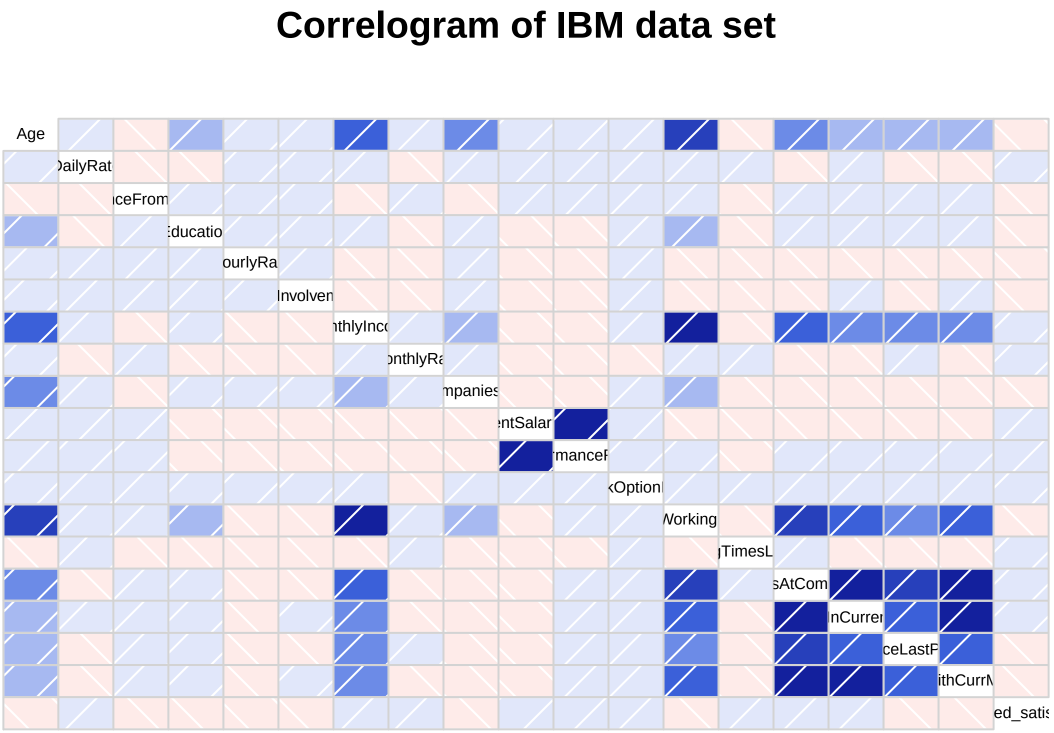
However, since JobSatisfaction (distrete) and WorkLifeBalance (discrete) are all subset of TotalSatisfaction, we can combined them together as a combined satisfaction (continues variable) and continue exploring which factors affect them through correlation test, in order to find out potential variables that indirectly affect attriton.

=======

Correlation Test with TotalSatisfaction relative factors

As stated in the previous analysis part, in this correlation part, we will continue exploring the relationship between combined satisfaction (JobSatisFaction & WorkLifeBalance) and other variables, seeking potential indirect fators contributed to employees' attrition choices.

First, we create a correlation table which creates a general idea on all correlations and we will then select potential variable for further correlation tests.


Correlogram of IBM data set

In the correlogram, the darker shade implies the higher correlation with each other. The last line indicates all correlations among combined satisfactions and other variables. The factors (with darker shade) need to be inspected and the potential resaons for selecting them are listed below.

Variables	Reasons
MonthlyIncome	Money might be the most direct reason for job satisfaction
PercentSalaryHike	Consider about furture development, people may be more happy with high Percent Salary Hike
PerformanceRating	Motivations for work might decrease due to lower performance rating
TrainingTimesLastYear	Longer training time could be helpful on employees' self-improvement
YearsInCurrentRole	Long time for repeating the same task might decrease the interest on work

1) Monthly Income

```
cor.test(IBM[,"MonthlyIncome"],IBM[,"combined_satisfaction"])
```

```
##
##  Pearson's product-moment correlation
##
## data:  IBM[, "MonthlyIncome"] and IBM[, "combined_satisfaction"]
## t = 0.40687, df = 1468, p-value = 0.6842
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04053079  0.06171263
## sample estimates:
##        cor
## 0.01061867
```

Table 5: Summary

Test_statistic	df	P_value	Correlation
0.4069	1468	0.6842	0.0106

As shown above, the correlation between Monthly Income and combined satisfaction are around 0.01, there is little relationship between them.However, the increase on monthly income doesn't influence that much as what we expected before. High P-value (around 0.68) indicates monthly income may not be a significant factor for combined satisfaction.

2) Percent Salary Hike

```
##
##  Pearson's product-moment correlation
##
## data:  IBM[, "PercentSalaryHike"] and IBM[, "combined_satisfaction"]
## t = 0.58278, df = 1468, p-value = 0.5601
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03594685  0.06628448
## sample estimates:
##        cor
## 0.01520856
```

Table 6: Summary

Test_statistic	df	P_value	Correlation
0.5828	1468	0.5601	0.0152

Similar to the previous test, it can be observed that combined satisafaction is possitively but weakly connected with distance from home. In this case, increasing on percent salary hike may have no impact on changing employee's satisfaction.

3) Performance Rating

```
##
##  Pearson's product-moment correlation
##
## data:  IBM[, "PerformanceRating"] and IBM[, "combined_satisfaction"]
## t = 0.12842, df = 1468, p-value = 0.8978
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04778400  0.05446977
## sample estimates:
```

```
##         cor
## 0.003351649
```

Table 7: Summary

Test_statistic	df	P_value	Correlation
0.1284	1468	0.8978	0.0034

With correlation value close to zero and 0.89 p-value, we can conclude that there is strongly evidence to reject performance rating as one of the variables for affecting attrition.

4) Training Times Last Year

```
##
##  Pearson's product-moment correlation
##
## data:  IBM[, "TrainingTimesLastYear"] and IBM[, "combined_satisfaction"]
## t = 0.39726, df = 1468, p-value = 0.6912
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.04078111  0.06146284
## sample estimates:
##        cor
## 0.01036796
```

Table 8: Summary

Test_statistic	df	P_value	Correlation
0.3973	1468	0.6912	0.0104

Similar as the privous factor, with 0.01 correlation value, increasing on traning time per year still does not have significant impact on combined satisfaction.

5) Years In Current Role

```
##
##  Pearson's product-moment correlation
##
## data:  IBM[, "YearsInCurrentRole"] and IBM[, "combined_satisfaction"]
## t = 0.9649, df = 1468, p-value = 0.3348
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02598514  0.07620513
## sample estimates:
##        cor
## 0.02517577
```

Table 9: Summary

Test_statistic	df	P_value	Correlation
0.9649	1468	0.3348	0.0252

Compared with other factors tested above, years in current role shows stronger positive correlation with combined satisfaction, which implies the longer employees working on the same role the higher satisfaction they have.

Conclusion for the correlation test

As we can see from the above correlation test, years in current role is more relative with combined satisafaction(JobSatisFaction & WorkLifeBalance) compared with others. This implies years in current role may be an indirect variable contributing to employees’ attrition choice. However, its correlation value is still not very significant. Therefore the linear regression might not give the a suitable approximation, it may be an approriate choice to apply another regression methods.

Analysis

Converting variables into Factor Variables

IBM$Education < -ordered(IBM$Education,levels=c(“1”,“2”,“3”,“4”,“5”)) IBM$JobSatisfaction < -ordered(IBM$JobSatisfaction,levels=c(“1”,“2”,“3”,“4”)) IBM$JobLevel < -ordered(IBM$JobLevel,levels=c(“1”,“2”,“3”,“4”) IBM$JobInvolvement < -ordered(IBM$JobInvolvement,levels=c(“1”,“2”,“3”,“4”)) IBM$JobLevel < -ordered(IBM$Joblevel,levels=c(“1”,“2”,“3”,“4”,“5”)) IBM$StockOptionLevel < -ordered(IBM$StockOptionLevel,levels=c(“ IBM$TrainingTimesLastYear < -ordered(IBM$TrainingTimesLastYear,levels=c(“0”,“1”,“2”,“3”,“4”,“5”,“6”)) IBM$RelationshipSatisfaction < -ordered(IBM$RelationshipSatisfaction,levels=c(“1”,“2”,“3”,“4”)) IBM$WorkLifeBalance < -ordered(IBM$WorkLifeBalance,levels=c(“1”,“2”,“3”,“4”))

Hypothesis Tests. Cannot Use t-tests as our independent variable is not continuous. Instead, we use regression analysis to determine whether or not the variable is statistically significant.

First, regressing Attrition on department. Not statistically significant

Regression Analysis.

The motivation for the following model is to provide the company with a clear indication of the factors significantly contributing to an employee leaving the company. We will then split the model into two: one for those factors the company can directly influence in order to decrease its attrition rate, the other outlaying those which are of a more personal nature and where the company has limited scope for interference. That is not to say that those findings have no value, though. The information uncovered here provides the company with a broad insight into the morale across the demographics of its workforce and could provide motivation for the design of programmes or events that subtly engage those with a higher probability of searching for another job.

Hence, we propose a model which treats Attrition as the dependent, or explanatory, variable. Our earlier analysis describes Attrition as a categorical variable and for this reason we used the method of logistic regression to build the model. This method provides the analyst with a percentage chance of an event occuring, where '1' is coded to mean there is a high chance that the employee will leave and '0' means the opposite. We use the particular method of logit smoothing to restrict our output between these levels. Incidentally, this also serves as a basis for a scale on which to compare the probability of individual employees leaving. For instance, if the company wanted to know the reasons behind people of or over a certain age leaving the company, then, controlling for age, it is possible to compare whether one employee is likely to leave over another based on their other key characteristics. From here, the company can direct effort into those areas to improve the chances of that employee choosing to stay at the company.

The process and logic of determining the model is such. First, we use a maximum-likelihood estimation (MLE) to quantify the effect of any particular variable on the probablity of an employee leaving the company. It must be noted that MLE assumes an aspymtotically normal distribution and thus invokes the use of the Wald statistic to test whether or not that particular variable's effect is statistically significant: our null hypothesis is always that the variable has no effect. We set a 95% confidence interval for the model. If the Wald statistic is deemed to be more extreme than this level, then the null hypothesis is rejected and we accept that the alternative hypothesis that there is a statistically significant relationship between those two variables; that the independent variable goes some way in explaining the probablity that an employee leaves. The z-value is then called and calculated to provide the probablilty that the null hypothesis is true in a value named the p-value. If the p-value is below 5% then we reject the null hypothesis.

z score to compute the p value mean = E(Y)=np var = np(1-p)

Wald test z-test

```r
att_1 <- glm(formula = Attrition ~ Department, data = IBM, family=binomial)
#z.test(x, y = DistanceFromHome, Attrition) #Can't figure out
summary(att_1)
```

```
##
## Call:
## glm(formula = Attrition ~ Department, family = binomial, data = IBM)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6797  -0.6501  -0.5458  -0.5458   1.9888
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -1.44692    0.32084  -4.510 6.49e-06 ***
## DepartmentResearch & Development -0.38175    0.33417  -1.142    0.253
## DepartmentSales                   0.09941    0.34152   0.291    0.771
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1298.6  on 1469  degrees of freedom
## Residual deviance: 1288.1  on 1467  degrees of freedom
## AIC: 1294.1
##
## Number of Fisher Scoring iterations: 4
```

Second, regressing Attrition on DistanceFromHome

att_2 <- glm(formula = Attrition ~ DistanceFromHome, data = IBM, family=binomial) summary(att_2)

Third, regressing Education on Attrition (NOT DONE YET)

att_3 <- glm(formula = Attrition ~ CONVERTEducation, data = IBM, family=binomial) summary(att_3)

Fourth, regressing Age on Attrition

att_4 <- glm(formula = Attrition ~ Age, data = IBM, family=binomial) summary(att_4)

Fifth, regressing Hourly Rate on Attrition. Not statistically significant

att_5 <- glm(formula = Attrition ~ HourlyRate, data = IBM, family=binomial) summary(att_1)

Sixth, regressing Monthly Income on Attrition.

att_6 <- glm(formula = Attrition ~ MonthlyIncome, data = IBM, family=binomial) summary(att_6)

Seventh, regressing YearsAtCompany on Attrition

att_7 <- glm(formula = Attrition ~ YearsAtCompany, data = IBM, family=binomial) summary(att_7)

Eighth, regressing PercentSalaryHike on Attrition

att_8 <- glm(formula = Attrition ~ PercentSalaryHike, data = IBM, family=binomial) summary(att_1)

Nine, regressing NumCompaniesWorked on Attrition. Not statistically significant.

att_9 <- glm(formula = Attrition ~ NumCompaniesWorked, data = IBM, family=binomial) summary(att_1)

Something.

tot_sat_11 <- glm(formula = Attrition ~ YearsAtCompany, data = IBM, family=binomial) summary(att_1)

Avi, this is the test regression

```
att_reg <- glm(formula = Attrition ~ DistanceFromHome + Age + MonthlyIncome + YearsAtCompany +
NumCompaniesWorked, data = IBM, family=binomial) summary(att_reg)
```

Reg Models Equation

Full Model

```
full_model_reg<-glm(Attrition~BusinessTravel+DailyRate+Department+DistanceFromHome+Education+EducationField+C
= IBM_Sqr, family= binomial)
```

```
options(digits=19) summary(full_model_reg)
```

Company_control_model

```
company_control_model<-glm(Attrition~BusinessTravel+DailyRate+Department+HourlyRate+JobInvolvement+JobLevel+J
= IBM_Sqr, family= binomial)
```

```
options(digits=19) summary(company_control_model)
```

out_of_control_model

```
out_of_control_model<-glm(Attrition~DistanceFromHome+Education+EducationField+Gender+MaritalStatus+NumCompa
= IBM_Sqr, family= binomial)
```

```
options(digits=5) summary(out_of_control_model)
```

Visualising Regression

Trying to visualize The generalized linear models created above

```
install.packages(“visreg”) library(“visreg”)
```

```
data(“IBM_Sqr”, package=“MASS”) fit <- glm(Attrition~DistanceFromHome+MonthlyIncome, data =
IBM_Sqr, family= binomial) visreg(fit, “DistanceFromHome”, xlab=“DistanceFromHome”, ylab=“probability
of Attrition”) visreg(fit, “MonthlyIncome”, xlab=“MonthlyIncome”, ylab=“probability of Attrition”)
```

Conclusion

Medina, E. (2012) Job satisfaction and employee turnover intention: What does organizational culture have to do with it. *Columbia University Academic Commons.*