DESAUTELS | McGill

# IBM Employee Attrition Analysis

## Final Project Report

Ramy Hammam
260888863

Shaher yar Jahangir
260924800

# Table of Contents

# Introduction

The challenge of employee retention is an everyday and ongoing issue for large companies like IBM. Today it is estimated that recruiting a new employee can cost up to $5,000 with around 42 days to fill the position. To minimize this cost, companies should identify subtle trends in employee behavior to maximize long-term employee stay.

The goal of this study is to understand if money is the main source of employee satisfaction and therefore retention. Otherwise, it is essential to determine the true factors that play an important role in determining the probability of an employee leaving. With this insightful information, IBM can better implement preventive measures to avert high employee turnover.

Different classification algorithms will be explored with "Attrition" being the dependent variable in order to identify the different interactions and how they influence employee attrition. In addition, the use of cluster analysis will support the creation of employee groups with different profiles and the identification of factors that determine the degree of an employee's state of happiness and satisfaction. This in return will allow companies like IBM to deploy tailor made efficient strategies to maximize the probability of retaining their employees.

# Data exploration

## Attrition Count

The dataset contains 1470 observations and 35 variables. Our goal is to identify the significant factors that lead to employee attrition at IBM. As seen in Figure 1, 237 employees have attrited which is equivalent to 16% of the total number of employees.

## Analysis of Monetary Variables

The variables associated with monetary factors are Monthly Income, Hourly Rate, Daily Rate, Monthly Rate. According to Figure 2, the employees who have attrited generally have lower monthly income and a lower daily rate. For Hourly Rate and Monthly Rate there is no discernable significant difference between the group of employees that have left and the employees that have not left the company.

## Overtime Analysis

The bar chart in Figure 3 indicates a relatively higher percentage of people working overtime in the group of employees that have attrited. 54% of the employees who quit work overtime versus 23% of employees who have not quit. As observed from the scatterplots in Figure 4, in the group of employees who have not attrited, the employees that work overtime are promoted faster than the ones that do not work overtime. The opposite can be noticed in the group of employees that have attrited. A pattern can be inferred that shows a group of employees leaving because they are not promoted although they work hard.

### Distance vs Work life & Business Travel

The boxplots in Figure 5 shows that the group of employees that have attrited and live close to their workplace have a high work life balance rating. It can be deduced that the distance from work affects the attrition variable. The bar graph in Figure 6 show that employees who travel more frequently on business trips are more likely to leave compared to those who do not.

### Attrition Rate within Departments

Figure 7 shows the attrition percentages in different departments. The highest turnover rate belongs to the sales department where 20.6% of employees in that department have attrited. The research and development and human resources department had a lower turnover rate.

### Marital Status, Income & Age Variables

The bar graph, as seen in Figure 8 shows that regardless of an employee's marital status there is a large number of employees who stay with the company and do not leave. Therefore, martial status is a weak predictor of attrition. Observing the scatter plot in Figure 9 the relationship between age and monthly income is linear across the different marital status. Single employees have more attitrion across age groups. The married employees with lower monthly income attrite more; whereas, divorced employees do not attrite much.

### Education Levels and Field of Education vs Attrition

The higher number of employees who have attrited are the ones that possess a bachelor's degree (Education level =3) and a master's degree (Education level =4) degrees in the Life Sciences and Medical fields as seen in Figure 10.

### Years of Experience

Referring to the Attrition trend with number of years of experience in Figure 11, Employees who have attrited have lower number of years of experience than the ones who do not attrit. The histogram in Figure 12 indicates that employees with less than 10 years of experience prefer to move to another company.

### Job Involvement and Job Satisfaction Analysis

Figure 13 shows the employees with high job involvement have higher attrition rates followed by medium job involved employees. Looking at the Job satisfaction graph in Figure 14, there is a visible increasing trend in the category of employees who do not attritate in contrast to employees who do attriate.

### Performance Analysis

The bar graph in Figure 15 of the performance variable against attrition, no visible difference can be seen between the performance rating and attrition levels. Both the groups of attrition level have similar percentages.

### Years at Company vs Monthly Income Analysis

As seen from Figure 16, the employees who do not attriate have higher monthly incomes than employees who do. As expected, there is a linear relationship between monthly income and the years at company variable.

### Key Findings

In conclusion, the factors most likely leading to employee attrition are: lower monthly income, working overtime, frequent traveling, long distance between home and workplace, high pressure to meet targets in the sales department and high job involvement.

## Model Selection & Methodology

### Classification Data Preprocessing

The dataset contains some variables that have zero variance or in other words, all the observations are the same. This can lead to spurious relationships that can affect the target variable. For example, the variables “EmployeeCount” “Over18” and “StandardHours” contain the same values for each row. To tackle this anomaly, the three columns were neglected from the dataset. The target variable was also converted from categorical values such as “Yes” or “No” to binary numbers where 1 represents “No” and 0 represents “Yes”. This was done since the target variable in logistic regression must be binary. In addition, mean encoding technique was utilized for categorical variables that have more than 5 levels on the training data. This encoding method explores the relation between similar categories, but the relations are bounded within the categories and target itself. Since the dataset is also rather small, too many variables created by dummification would prevent the algorithm to return good performances as there is insufficient data to go over the different scenarios. Mean encoding mitigates this performance issue.

## Clustering Data Preprocessing

The initial dataset was refreshed when utilized for the clustering tasks. Similarly, to classification, variables with zero variance were dropped. The remaining variables were normalized since clustering works with distance metrics. This means that objects will be grouped together given a measure of similarity. Normalization ensures the grouping and avoids the creation of artificial dissimilarities across observations when none exist.

## Model Selection and Methodology

When crafting an optimal robust model, different classification and clustering techniques were taken into consideration to study the effect on employee attrition. The two classification methods chosen were Logistic Regression and XGBoosting. Logistic regression classifier proved to be an obvious option due to its easy interpretability and strong performance. Boosting was utilized as it is considered one of the best classification methods that leverages the power of simple trees which grow sequentially and learn from each other to achieve the highest accuracy. The comparison will be based on whichever model outputs the highest F1 score rather than just accuracy since the F1 score accounts for recall and precision metrics. The dataset was then split with 70 – 30 training to test ratio with 2-fold cross-validation utilization.

The objective for applying logistic regression was not to enhance our model but to explore the variables that positively or negatively affect the target variable. A summary of the logistic regression model was used to determine the significance level of each predictor. Variables that are associated to the likelihood that the relationship with the target variable is caused by something other than pure chance were separated into two tables based on their significance ($P<0.05$).

Additionally, the increase of probability of an employee being retained at the company was explored with the effect_perc column in Table 1. This was calculated by exponentiating the variable coefficients and finding the compliment to determine the percentage change.

Next, K-means clustering was performed using only the significant variables identified from logistic regression. The selected value of k clusters was three clusters based on the facilitation of model interpretation and limited size of the dataset. A PCA analysis will perform a dimensionality reduction expressed by the two axes of a biplot and a correlation plot which is a suitable tool to describe the different clusters characteristics.

# Results

## Logistic Regression vs Boosting

Logistic Regression model returned the more superior result against XGBoost with an F1 score of 93.9%. It is crucial to check for the statistical significance of parameter coefficients to ensure that any conclusion made on the relationship of a covariate on the dependent variable is reliable. As seen in Table 1, the most influential factors in whether or not an employee will attrite depends on if an employee: works overtime, is satisfied with the work environment, is highly involved in his job, and travels frequently for business. It is worth mentioning that monthly income of an employee is not a realistic determinant of attrition.  Moreover, the variables present in Table 1 also display a high sense of impactfulness (effect_perc) on the decision of attrition when compared to the least influential predictors.

| . | Estimate | Std. Error | z value | Pr(>\|z\|) | effect | effect_perc. |
|---|---|---|---|---|---|---|
| JobInvolvement | 0.6013958 | 0.1470408 | 4.089993 | 0.0000431 | 1.8246640 | 82.46640 |
| EnvironmentSatisfaction | 0.4972892 | 0.1034113 | 4.808846 | 0.0000015 | 1.6442580 | 64.42580 |
| WorkLifeBalance | 0.4750352 | 0.1562287 | 3.040640 | 0.0023608 | 1.6080708 | 60.80708 |
| JobSatisfaction | 0.3549701 | 0.1001114 | 3.545752 | 0.0003915 | 1.4261380 | 42.61380 |
| RelationshipSatisfaction | 0.3361028 | 0.1024364 | 3.281087 | 0.0010341 | 1.3994829 | 39.94829 |
| TrainingTimesLastYear | 0.2680209 | 0.0904028 | 2.964742 | 0.0030294 | 1.3073745 | 30.73745 |
| YearsInCurrentRole | 0.1610215 | 0.0554865 | 2.901992 | 0.0037080 | 1.1747102 | 17.47102 |
| YearsWithCurrManager | 0.1232788 | 0.0618080 | 1.994544 | 0.0460927 | 1.1311998 | 13.11998 |
| Ã¯..Age | 0.0535188 | 0.0180431 | 2.966157 | 0.0030155 | 1.0549768 | 5.49768 |
| DistanceFromHome | -0.0513117 | 0.0131060 | -3.915138 | 0.0000904 | 0.9499825 | -5.00175 |
| NumCompaniesWorked | -0.1901154 | 0.0469516 | -4.049181 | 0.0000514 | 0.8268637 | -17.31363 |
| YearsSinceLastPromotion | -0.1906597 | 0.0526712 | -3.619812 | 0.0002948 | 0.8264138 | -17.35862 |
| MaritalStatusSingle | -1.1092951 | 0.4354907 | -2.547230 | 0.0108582 | 0.3297914 | -67.02086 |
| BusinessTravelTravel_Rarely | -1.1153817 | 0.4820656 | -2.313755 | 0.0206812 | 0.3277901 | -67.22099 |
| OverTimeYes | -1.9810368 | 0.2379816 | -8.324326 | 0.0000000 | 0.1379262 | -86.20738 |
| BusinessTravelTravel_Frequently | -2.1153353 | 0.5190707 | -4.075236 | 0.0000460 | 0.1205928 | -87.94072 |

*Table 1: List of Significant Variables*

For example, the involvement level with the job, gives the strongest impact on employee retention where the increase of every unit in Job Involvement is followed by an increase in the retention probability of 82%. Another example is the different department an employee works in, which states that the probability of retention is only 46% if the employee works in sales as seen in Table 2.

## K-means & PCA Clustering

The correlation plot in Figure 17 supports the interpretation of the PCA axis, which represents the dimensions. The cos2 measure shows the importance of a principal component for a given observation. Dimension 1 (x-axis) has a strong correlation with the significant variables: YearsCurrentRole, YearsCurrentManager & YearsSinceLastPrmotion. Hence, the further to the right an employee is positioned in that plot, the stronger it will correlate to those variables. Subsequently, the employees will have a negative correlation with WorkLifeBalance variable. In addition, the second dimension (y-axis) states that the further to the upper half an employee is

positioned, the stronger will the correlation be towards Business_Travel_Rarely. Also, employees correlate negatively against Business_Travel_Frequency as expected.

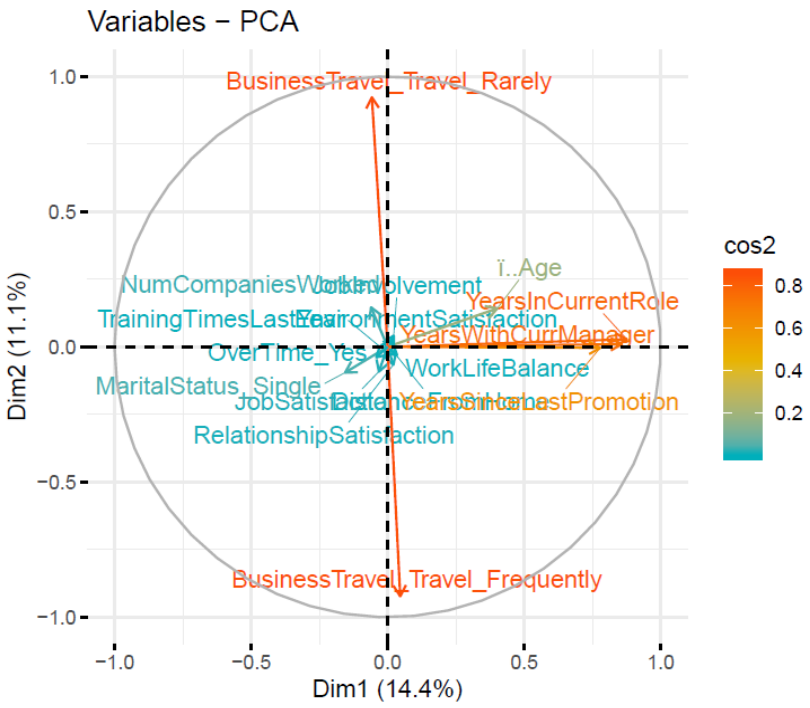


*Figure 17: Principal Component Analysis (PCA) Plot*

## Clustering Analysis

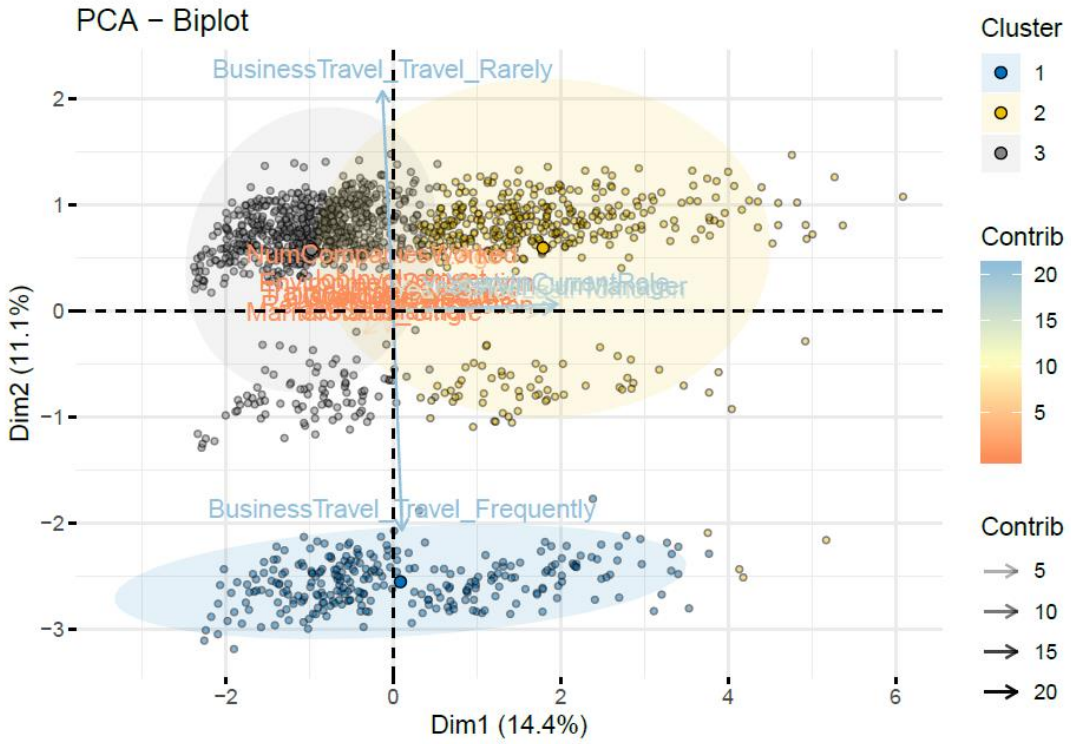


*Figure 18: Clustering Plot*

After having performed the K-Means clustering method, where 3 clusters have been accounted for, cluster 1 as seen in Figure 18 (lower half of the correlation biplot) characterizes employee groups that have medium levels of attrition, age and monthly income. In addition, they do not display any particular high levels of job satisfaction, but in fact have relative low levels of job dissatisfaction. Hence, one could assume that those are more experienced employees that are not heavily overworked and tend to be remunerated, travel more and have worked more often at previous companies (as seen in Table 4, Dim.2 column). However, the consequence here is that those employees that travel more frequently have a higher probability of leaving the company of 87%

as opposed to individuals that do not travel. Traveling frequently and traveling rarely strongly correlate with attrition. Having a proper balance in traveling is necessary to lower the chances of employee attrition.

Cluster 2 has the individuals with the highest level of seniority at the company with the largest monthly income, that have been working the longest time at the current role, with the current manager and have waited a long time for a promotion. In other words, these employees, that are positioned more to the further of the biplot, express the seniority level which can be a positive indicator, 17% and 13% respectively for every additional year or a negative indicator. This reason being, employees waiting for too long for a promotion might be willing to leave the company with chances of 17% for every extra year of waiting. Hence, one could state that in general employees in cluster 2 even though older than employees from cluster 1 and 3, have levels of attrition that are lower than the other groups. However, senior employee's level of job and environment satisfaction are not the highest, yet they do travel much less, which has been proved to be a strong factor contributing attrition. They also have on average a higher monthly income than the rest. Thus, a connection of low attrition levels combined with higher monthly income can be deduced, which is not necessarily followed by higher and better levels of job satisfaction but just with better conditions overall such as not working as much overtime.

Next, cluster 3 contains employees that are more heavily exploited in terms of working overtime, having lower levels of monthly income, traveling rarely and being on average younger than the employees in other groups. For the factor of not travelling frequently the chances that these individuals leave the company increases by 67% as shown by Table 1 Hence, these are most likely the employees that have recently started since they correlate negatively with variables related to seniority such as working years at the company or waiting for a promotion. As a result, individuals

from cluster 3 are employees that do not have a positive standing at the company in terms of career prospective. The longer the time spent at the company and the less they get to travel the more demotivated they will feel and the higher are the chances of attrition as supported by Table 1. In conclusion, they have the highest levels of job dissatisfaction at the lower scales and have, compared to the other clusters, the highest levels for probability of attrition.

## Predictions and Conclusion

The following connections have been made after having analyzed relationship between attrition and various variables through the logistic regression model and the k-means cluster analysis. First, retention is a result of good remuneration and low levels of employee exploitation in terms of low extra working hours and medium levels of traveling. It is not possible to state by how much the chances of retention increase the more the salary is increased since the parameter estimate is not statistically significant, meaning that no reliable inference can be made. Thus, through descriptive analysis and dismantling the clusters, we can graphically explain the relationship. However, if an employee travels frequently the chances for him to leave the company are 87% higher than for non-travelers.

Moreover, job satisfaction does not seem to be a consequence of good remuneration as clusters with the highest level of monthly income did not reflect that. Nonetheless, it has a positive impact on attrition, where an additional increase in the scale of job satisfaction increases the chances of retention by 42%. Hence, the wellbeing of an employee is clearly important to ensure that he remains at the company. The logistic regression model clearly stated that the top 2 factors motivating the employee to stay at the company are variables related to the employee's well-being, such as work life balance and job satisfaction. For the first mentioned factor, an increase of one

unit in the scale of work life balance is followed by an increase of 60% in the probability of retention. The importance of the employee's wellbeing should be a priority for the company.

Therefore, attrition is strongly correlated to whether an employee feels well-being which is supported by the logistic regression probability outputs. The better an employee feels the more satisfied he will be, the better the work life balance and the higher are the chances for him to not quit.

However, job satisfaction does not appear to be present simultaneously with high levels of income as the cluster analysis suggests. In fact, individuals with higher levels of income tend to have medium to low levels of Job satisfaction. Moreover, the chances of attrition are counteracted by higher salaries. Thus, the final statement would be that a company needs to either ensure a good work life balance with little stress upon the employees or ensure that salaries are high enough to mitigate the stress. Since the allocation of salary is dependent on the level of seniority, the short-term solution to stop the levels of attrition would be to lower the burden of too many working hours and ensure a better working environment for junior employees.

# Appendix

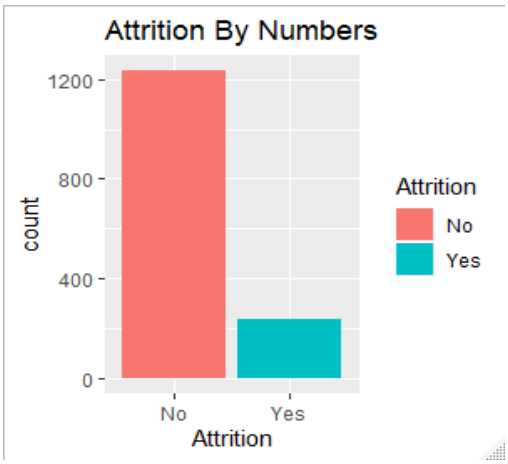


*Figure 1: Bar Graph of Attrition by Numbers*

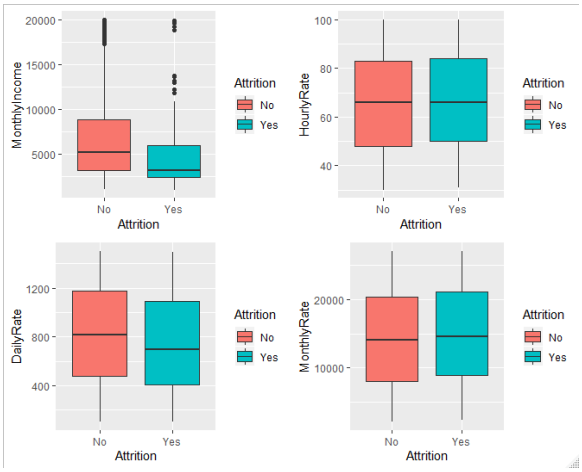


*Figure 2: Boxplot Analysis of Monetary Variables*

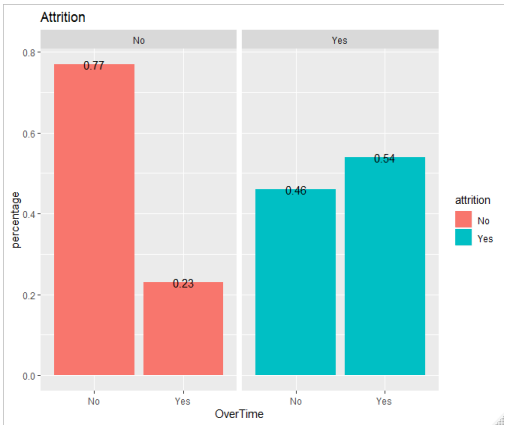


*Figure 3: Overtime Analysis Bar Graph*

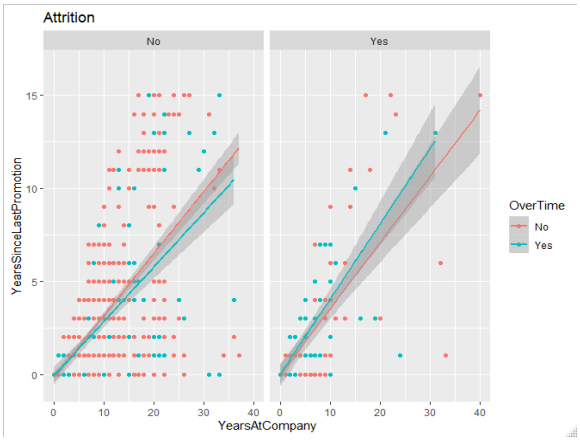


*Figure 4: Overtime Analysis Scatterplot*

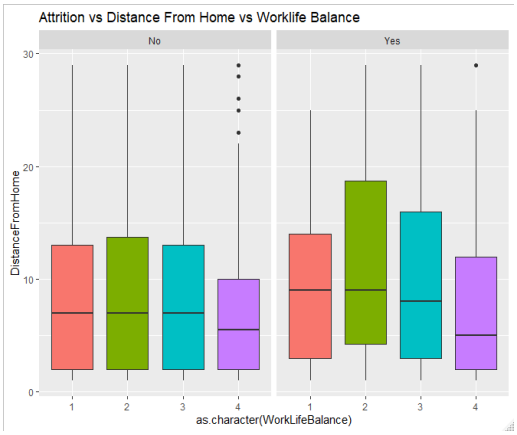


*Figure 5: Boxplot of Attrition % vs Distance vs Work life*

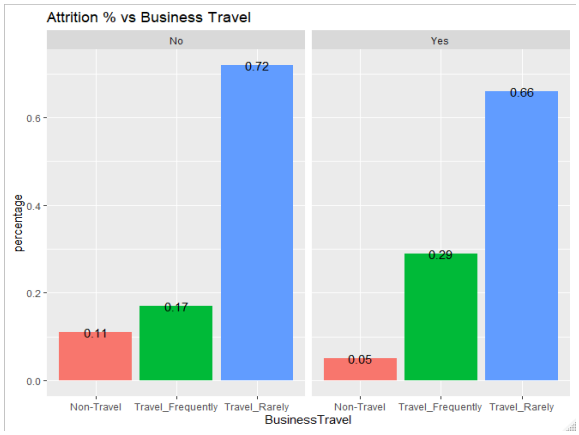


*Figure 6: Bar Graph of Attrition % vs Business Travel*

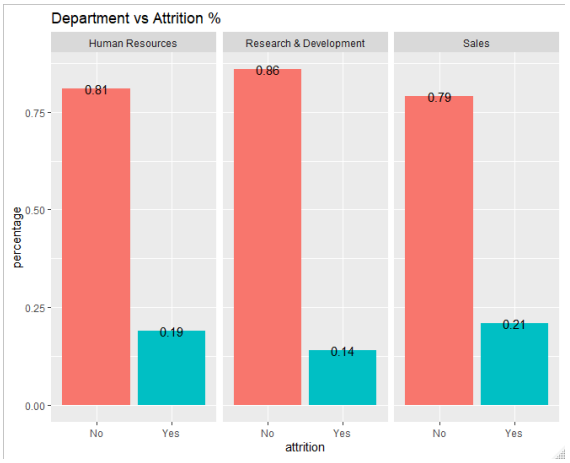


*Figure 7: Department vs Attrition %*

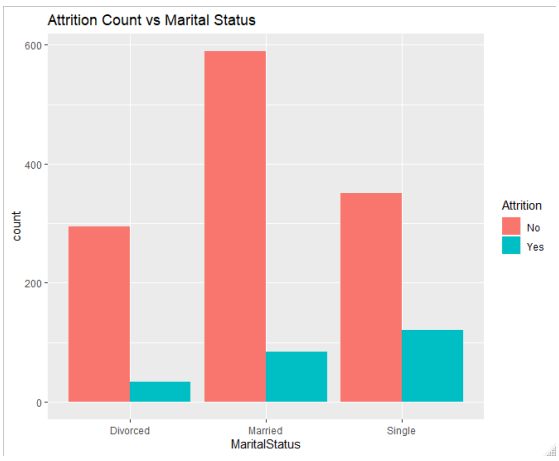


*Figure 8: Attrition % vs Business Travel*

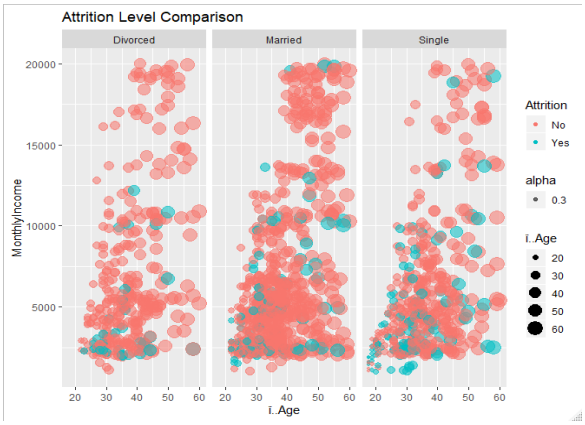


*Figure 9: Attrition Level Comparison with Age & Monthly Income*

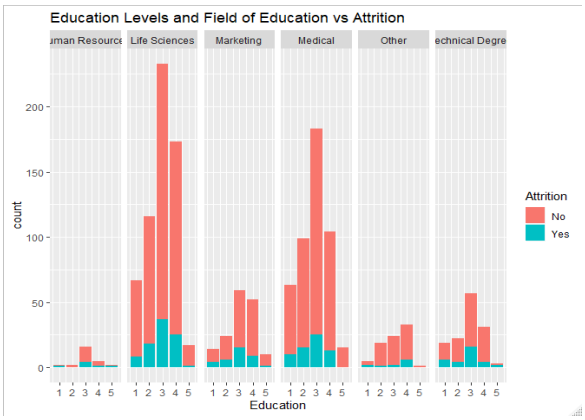


*Figure 10: Education Levels and Field of Education vs Attrition*

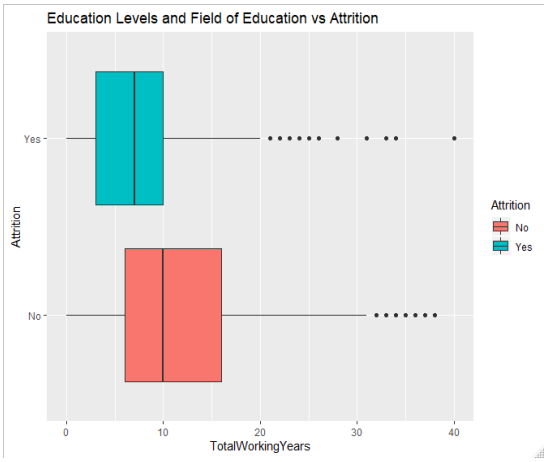


*Figure 11: Boxplot of Edu Levels and Field of Edu vs Attrition*


*Figure 12: Histogram of Experience/ Attrition Yes, No*

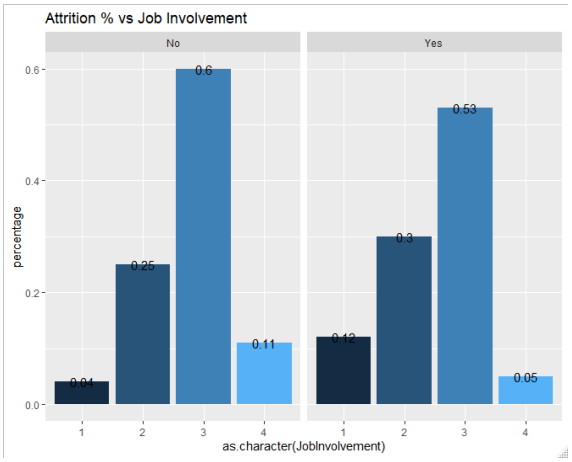


*Figure 13: Bar Graph of Attrition % vs Job Involvement*

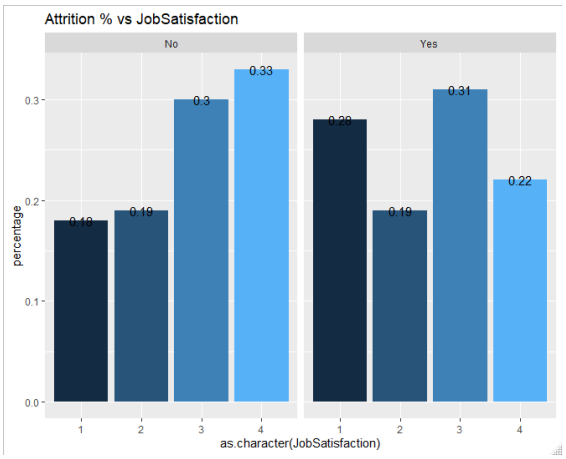


*Figure 14: Bar Graph of Attrition % vs Job Satisfaction*

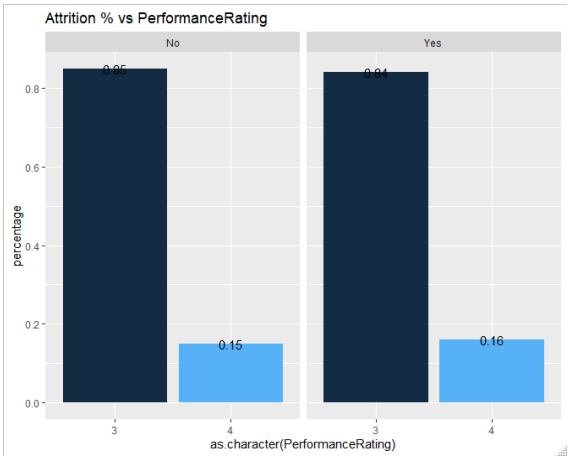


*Figure 15: Bar Graph of Attrition % vs Performance Rating*

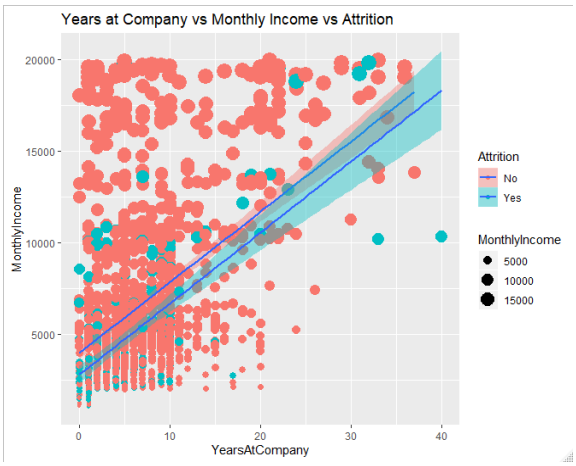


*Figure 16: Scatterplot of Years vs Monthly Income vs Attrition*

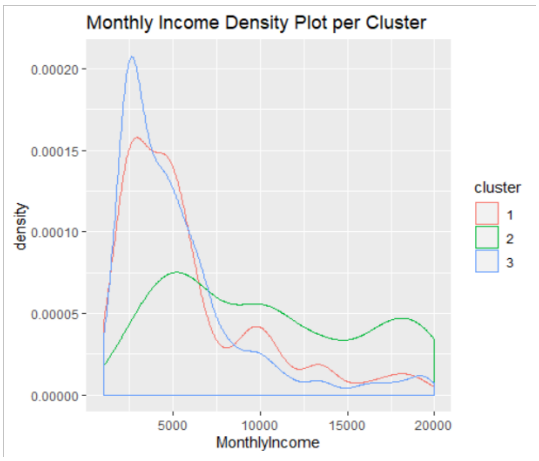


*Figure 19: Monthly Income Density Plot per Cluster*


*Figure 20: Attrition Density Plot per Cluster*


*Figure 21: Age Density Plot per Cluster*

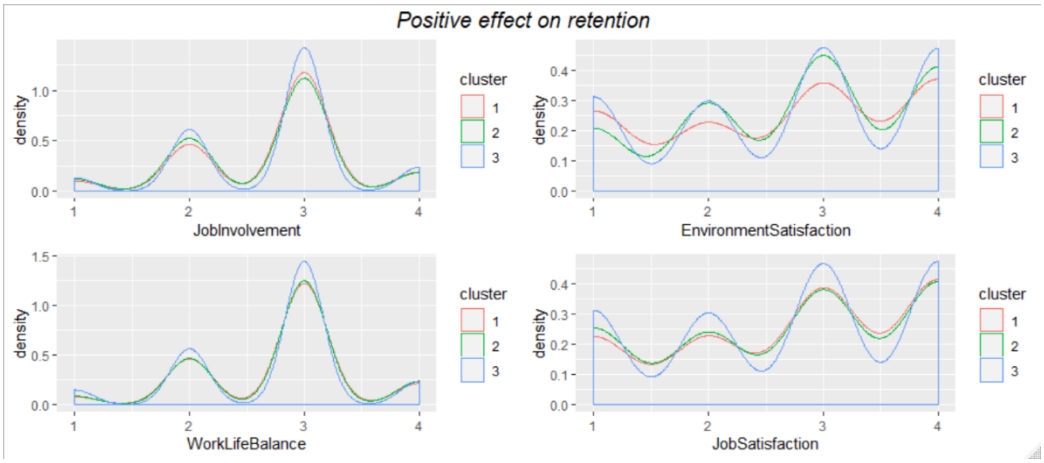


*Figure 22: Positive Effect on Retention*

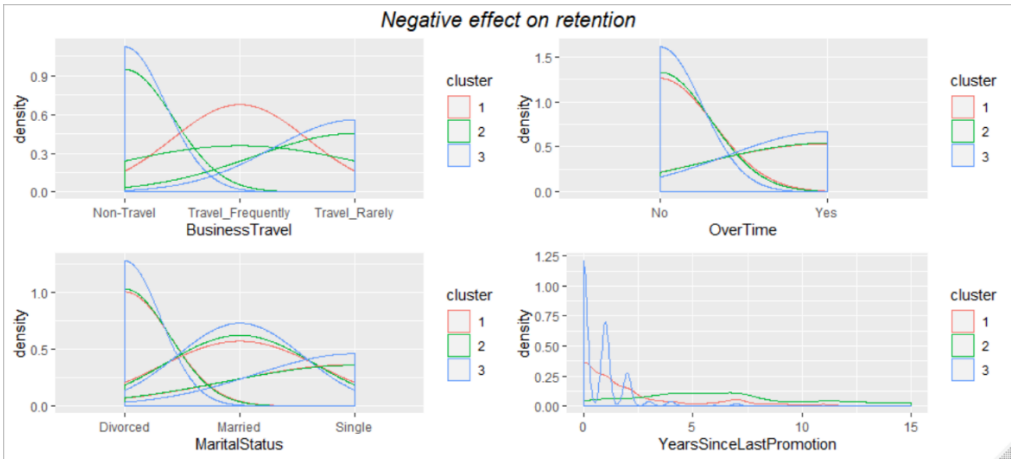


*Figure 23: Negative Effect on Retention*

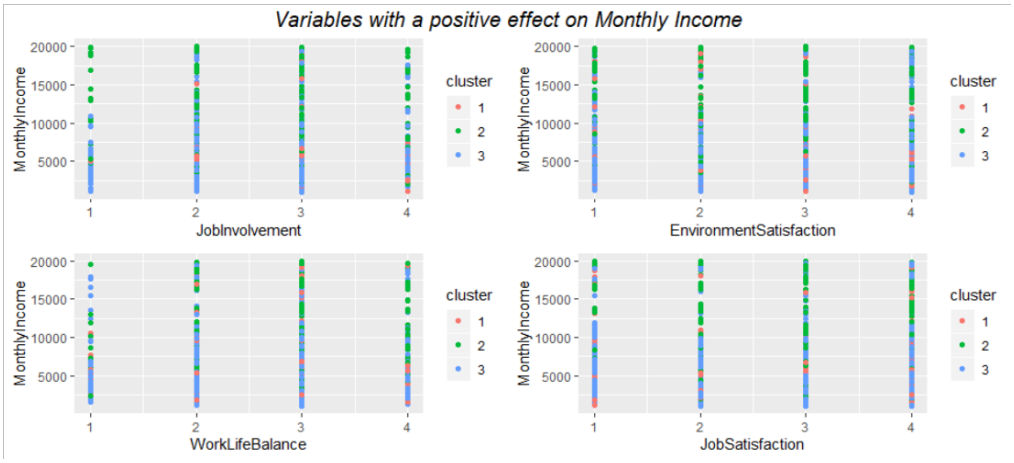


*Figure 24: Variable with a Positive Effect on Monthly Income*

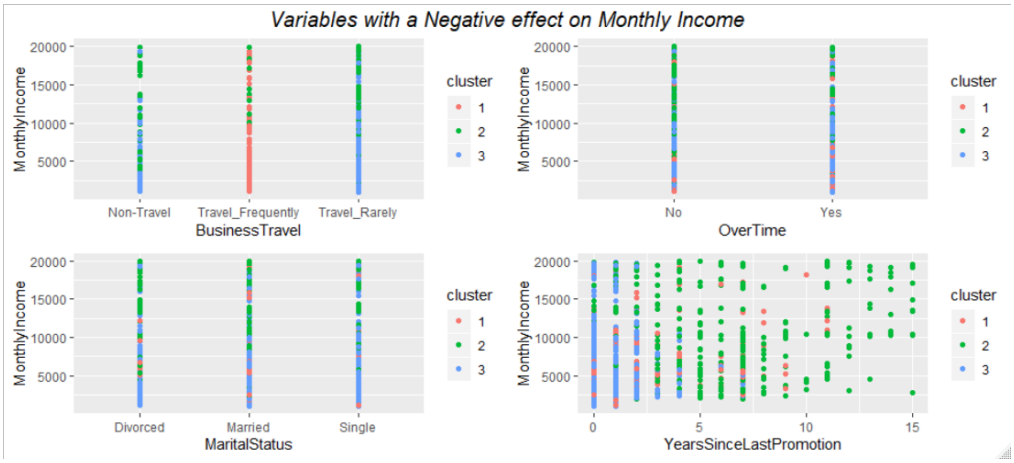


*Figure 25: Variable with a Negative Effect on Monthly Income*

| . | Estimate | Std. Error | z value | Pr(>\|z\|) | effect | effect_perc. |
|---|---|---|---|---|---|---|
| DepartmentSales | 0.3798677 | 0.5363815 | 0.7082042 | 0.4788185 | 1.4620911 | 46.2091079 |
| StockOptionLevel | 0.2688658 | 0.2046892 | 1.3135318 | 0.1890038 | 1.3084795 | 30.8479474 |
| `DepartmentResearch & Development` | 0.0891160 | 0.5331137 | 0.1671613 | 0.8672432 | 1.0932074 | 9.3207417 |
| TotalWorkingYears | 0.0548777 | 0.0383481 | 1.4310386 | 0.1524191 | 1.0564114 | 5.6411357 |
| PercentSalaryHike | 0.0373513 | 0.0484309 | 0.7712284 | 0.4405716 | 1.0380576 | 3.8057584 |
| DailyRate | 0.0001485 | 0.0002733 | 0.5432486 | 0.5869587 | 1.0001485 | 0.0148475 |
| EmployeeNumber | 0.0001317 | 0.0001878 | 0.7010149 | 0.4832937 | 1.0001317 | 0.0131681 |
| MonthlyIncome | 0.0000354 | 0.0000904 | 0.3916059 | 0.6953495 | 1.0000354 | 0.0035391 |
| MonthlyRate | 0.0000055 | 0.0000154 | 0.3545058 | 0.7229598 | 1.0000055 | 0.0005471 |
| HourlyRate | -0.0062617 | 0.0054992 | -1.1386515 | 0.2548485 | 0.9937578 | -0.6242164 |
| PerformanceRating | -0.0221844 | 0.5043563 | -0.0439856 | 0.9649159 | 0.9780598 | -2.1940162 |
| MaritalStatusMarried | -0.0568058 | 0.3381592 | -0.1679855 | 0.8665947 | 0.9447775 | -5.5222496 |
| YearsAtCompany | -0.0731341 | 0.0494572 | -1.4787357 | 0.1392110 | 0.9294762 | -7.0523813 |
| Education | -0.1083596 | 0.1109752 | -0.9764305 | 0.3288512 | 0.8973049 | -10.2695132 |
| GenderMale | -0.1646912 | 0.2231636 | -0.7379841 | 0.4605241 | 0.8481556 | -15.1844409 |
| JobLevel | -0.2869217 | 0.3965447 | -0.7235545 | 0.4693393 | 0.7505705 | -24.9429486 |

*Table 2: Non-Significant Variables from Logistic Regression*

| | Dim.1 | Dim.2 |
|---|---|---|
| Ã¯..Age | 0.4016908 | 0.1421994 |
| DistanceFromHome | 0.0320969 | -0.0315192 |
| EnvironmentSatisfaction | 0.0174669 | 0.0249374 |
| JobInvolvement | 0.0222943 | 0.0412683 |
| JobSatisfaction | -0.0317124 | -0.0913495 |
| NumCompaniesWorked | -0.0605287 | 0.1459935 |
| RelationshipSatisfaction | 0.0211711 | -0.0660495 |
| TrainingTimesLastYear | -0.0117367 | -0.0131333 |
| WorkLifeBalance | 0.0342579 | -0.0301369 |
| YearsInCurrentRole | 0.8762270 | 0.0279303 |
| YearsSinceLastPromotion | 0.7786564 | -0.0038144 |
| YearsWithCurrManager | 0.8576707 | 0.0101796 |
| BusinessTravel_Travel_Frequently | 0.0463297 | -0.9225542 |
| BusinessTravel_Travel_Rarely | -0.0575391 | 0.9229037 |
| MaritalStatus_Single | -0.1552720 | -0.0950481 |
| OverTime_Yes | -0.0412988 | -0.0388592 |

*Table 3: PCA Dimensions*