Analysis on employee attrition of a company

Shivangi Sharma

2022-10-05

Introduction

Employee attrition is defined as an upredictable and uncontrollable reduction in an organization's workforce. It could be due to several reasons like resignation, retirement or death. It affects the organization's performance, productivity and profitability in negative ways. The attrition rate of a company is a metric that provides insights into how well the said company is retaining its employees. In this analysis, I have studied a company's dataset on employee attrition to determine the causes that lead to this phenomenon and the relationship between those variables.

Based on a recent article that I read, one of the biggest challenges that a company is facing is retaining employees. That inspired me to carryout this analysis. I have picked out a dataset from Kaggle that contains data of around 1500 employees of a fictitious company. The link to the dataset can be found below:

https://www.kaggle.com/datasets/whenamancodes/hr-employee-attrition

I started by loading the dataset into R:

R code:

```
# Load the dataset into R
library(readr)
df <- read.csv("HREmployeeAttrition.csv")
```

Data Refinement

The dataset is created by IBM data scientists and it contains data of 1500 employees of an organization about their age, income, education, job role and other general information. It also contains data about which employees quit the company, their overall satisfaction level, work life balanace, years at company, salary hike percent and more. For my analysis, I have removed unwanted data from the dataset such as employee number, standard hours of work, whether the employee is over 18 and their job level. I have carried out further analysis with this updated dataset.

R code:

```
# Data cleaning
df <- subset(df, select = -c(EmployeeCount, EmployeeNumber, JobLevel, Over18, StandardHours))
head(df)
```

Which age group saw the highest attrition rate in the company?

Assuming that the company employs people over 18 years and the retirement age is 60 years. I divided the age into different intervals and plotted it in a stacked bar chart against total employees in that age group. I chose this graph type since it is easier for comparison. Each bar represents the percentage of employees that quit in that particular age group. It is evident from the graph that most of the employees who quit belong to the age group 18-24, as expected. This could be because young employees have more freedom and time to explore career opportunities or to go back to school to change career fields as opposed to older employees.

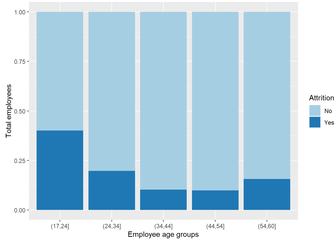
R code:

```
# Creating age group intervals
df$Attrition <- as.factor(df$Attrition)
df$age.grps <- cut(df$Age, c(17, 24, 34, 44, 54, 60))

# Graph 1: Age groups vs employee attrition
library(ggplot2)
library(RColorBrewer)

pl1 <- ggplot(df, aes(fill=Attrition, y=PercentSalaryHike, x=age.grps)) + geom_bar(position="fill", stat="identity") + labs(x="Employee age groups", y="Total employees")

pl1 + scale_fill_brewer(palette = "Paired")
```



Which department of the company sees the highest attrition rate?

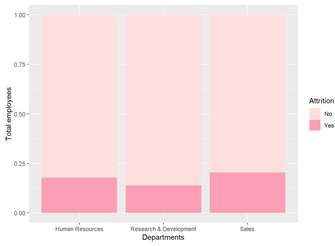
The company has three main departments: Human Resources, Research & Development and Sales. The stacked barplot depicts the department-wise attrition rates. The sales department has the highest attrition rate of all implying that employee turnover is high in this department. This could be due to low job satisfaction level or low income standards in the department.

R code:

```
# Classifying company departments as unique
unique(df$Department)

# Graph 2: Department wise employee attrition
pl2 <- ggplot(df, aes(fill=Attrition, x=Department, y=PercentSalaryHike)) + geom_bar(position = "fill", stat = "identity") + labs(x="Departments", y="Total employees")

pl2 + scale_fill_brewer(palette = "RdPu")
```



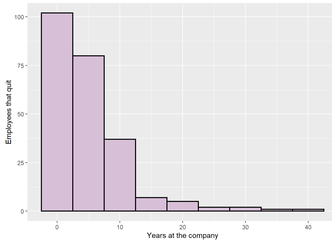
How long do employees stick around before they quit?

Companies want to increase their retention rate by hiring employees who will stay with them long term. This demonstrates a positive and stable environment and reduces hiring costs. To analyse the number of years most employees who quit worked for, I have created a separate data frame (df1) that contains data of only the employees who quit the company. I have shown the relationship in a histogram. I chose to depict this distribution in a histogram as it communicates the frequency of each interval effectively. It is evident from the histogram that most employees who quit, worked for less than 5 years in the company. The distribution suggests that the longer you keep working with the same company, the less likely you are to quit.

R code:

```
# Create new data frame with data of only the employees that quit
df1 <- df[df$Attrition == "Yes",]
head(df1)

# Graph 3: Histogram
ggplot(df1, aes(x=YearsAtCompany)) + geom_histogram(binwidth = 5, color = "black", lwd = 0.75, linetype = 1, fill = "thistle", position="identity") + labs(x="Years at the company", y="Employees that quit")
```



Is there any correlation between employees that quit and their payscale?

One of the major reasons why employees quit their jobs is payscale. Employees who believe that they are not being compensated fairly for the work that they are doing seek out better opportunities outside of the organization. To verify this relationship, I have plotted the company employees' income against their age based on whether they have quit the company. I have depicted it in a scatter plot as I wanted to show a correlation between the two variables. The blue dots depict the employees that have quit against the employees that are still part of the company depicted in red. As we can see, most of the blue dots are concentrated around the bottom left of the plot suggesting that most of the employees who quit had lower income compared to the ones who stayed. Another conclusion that we can draw from this plot is that the blue dots are more scattered on the right side, depicting low attrition among older workers who are paid more.

R code:

```
# Graph 4: Income vs attrition
ggplot(df, aes(x = Age, y = MonthlyIncome, color = Attrition)) + geom_point(alpha=0.5, position = position_jitter()) + theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(), panel.background = element_blank(), axis.line = element_line(colour = "black")) + labs(y = "Monthly Income")
```

What's the distribution of income between education levels in the company?

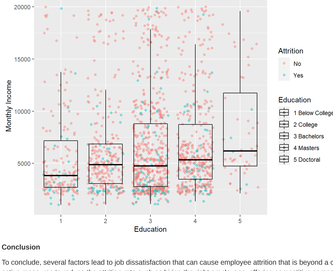
More education leads to better prospects for earnings and employment. This company has employees at various level of education and it is expected that higher the qualification, higher the pay. For this depiction, I have used boxplots and combined it with a scatter plot as well. The boxplots show the spread of the income. The points add another dimension to the data by showing the concentration of the number of employees at each education level. From the boxplot, we can verify that the highest level of education (Doctoral), does have a higher median pay. Moreover, most of the blue dots representing employees that quit fall below the first quartile, suggesting that employees who quit were getting paid lower than 75% of the employees with the same qualification.

R code:

```
# Graph 5: Income vs education level
df$Education <- as.factor(df$Education)

pl3 <- ggplot(df, aes(x = Education, y = MonthlyIncome, color = Attrition, fill = Education)) + geom_point(alpha = 0.4, position = "jitter") + geom_boxplot(alpha = 0, color = "black") + labs(y="Monthly Income")

pl3 + scale_fill_discrete(name="Education", labels=c("1 Below College","2 College","3 Bachelors", "4 Masters", "5 Doctoral"))
```



Conclusion

To conclude, several factors lead to job dissatisfaction that can cause employee attrition that is beyond a company's control. Companies can take active measures to reduce the attrition rate such as hiring the right employees, offering competitive compensation and benefits, offer career progression opportunities and promoting employee engagement. All these measures help increase an organization's productivity and overall business performance.