Unveiling the Secrets Behind Employee Attrition at IBM: A Data-Driven Approach

The Motivation Behind the Project

In the competitive landscape of modern businesses, retaining talent is critical. IBM, a global tech giant, has faced a surge in employee attrition recently. As a People Data Analyst intern, I embarked on a mission to uncover the factors contributing to this trend, using R and advanced statistical techniques.

Why This Analysis Matters

Understanding the root causes of employee turnover can save organizations substantial costs associated with hiring and training new employees. This article delves into the intricacies of attrition at IBM, revealing insights that could help not only IBM but also other companies facing similar challenges. You'll learn about the data analysis process, key findings, and actionable takeaways to mitigate attrition.

Exploring the Dataset

The dataset used for this analysis is a well-known HR dataset created by IBM data scientists. It contains 1470 rows representing employees and 35 columns with various attributes, including demographics, job-related information, and the critical "Attrition" column. The dataset is available for public use and can be found here.

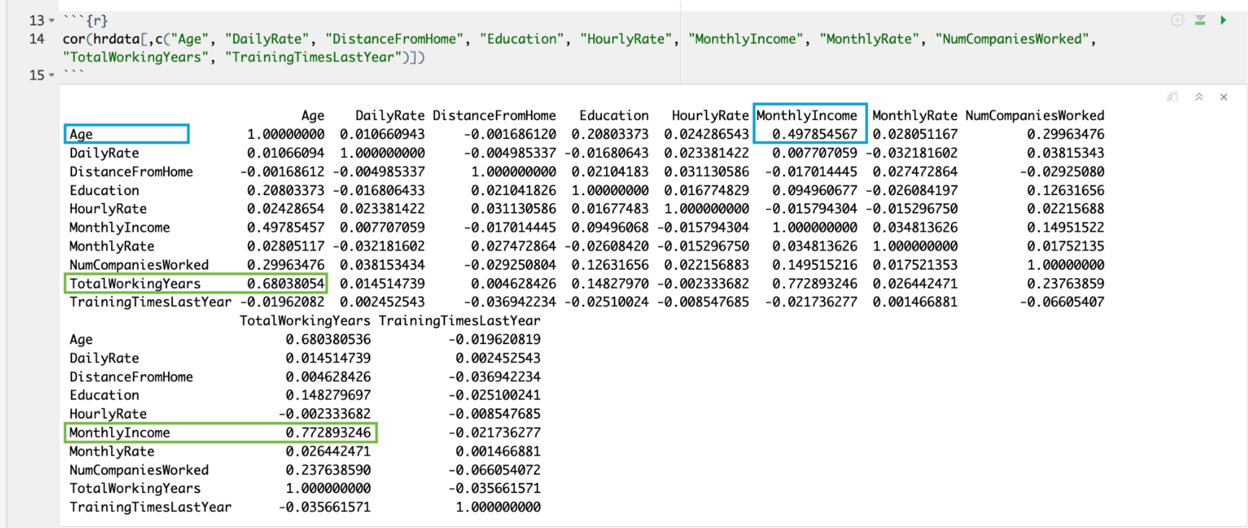
In-depth Data Analysis

Preparing and Cleaning the Data

First, I imported the dataset into R and conducted an initial exploration to understand the structure and contents. The key variable, "Attrition," was analyzed to differentiate between employees who stayed and those who left.

Correlation Matrix Insights

A correlation matrix was generated for important demographic and job-related variables such as Age, Monthly Income, and Total Working Years.



~~~
```{r}
cor(hrdata[,c("Age", "DailyRate", "DistanceFromHome", "Education", "HourlyRate", "MonthlyIncome", "MonthlyRate", "NumCompaniesWorked",
"TotalWorkingYears", "TrainingTimesLastYear")])
```
~~~

```
                              Age    DailyRate DistanceFromHome   Education   HourlyRate MonthlyIncome  MonthlyRate NumCompaniesWorked
Age                    1.00000000  0.010660943     -0.001686120  0.20803373  0.024286543   0.497854567  0.028051167         0.29963476
DailyRate              0.01066094  1.000000000     -0.004985337 -0.01680643  0.023381422   0.007707059 -0.032181602         0.03815343
DistanceFromHome      -0.00168612 -0.004985337      1.000000000  0.02104183  0.031130586  -0.017014445  0.027472864        -0.02925080
Education              0.20803373 -0.016806433      0.021041826  1.00000000  0.016774829   0.094960677 -0.026084197         0.12631656
HourlyRate             0.02428654  0.023381422      0.031130586  0.01677483  1.000000000  -0.015794304 -0.015296750         0.02215688
MonthlyIncome          0.49785457  0.007707059     -0.017014445  0.09496068 -0.015794304   1.000000000  0.034813626         0.14951522
MonthlyRate            0.02805117 -0.032181602      0.027472864 -0.02608420 -0.015296750   0.034813626  1.000000000         0.01752135
NumCompaniesWorked     0.29963476  0.038153434     -0.029250804  0.12631656  0.022156883   0.149515216  0.017521353         1.00000000
TotalWorkingYears      0.68038054  0.014514739      0.004628426  0.14827970 -0.002333682   0.772893246  0.026442471         0.23763859
TrainingTimesLastYear -0.01962082  0.002452543     -0.036942234 -0.02510024 -0.008547685  -0.021736277  0.001466881        -0.06605407
                      TotalWorkingYears TrainingTimesLastYear
Age                         0.680380536          -0.019620819
DailyRate                   0.014514739           0.002452543
DistanceFromHome            0.004628426          -0.036942234
Education                   0.148279697          -0.025100241
HourlyRate                 -0.002333682          -0.008547685
MonthlyIncome               0.772893246          -0.021736277
MonthlyRate                 0.026442471           0.001466881
NumCompaniesWorked          0.237638590          -0.066054072
TotalWorkingYears           1.000000000          -0.035661571
TrainingTimesLastYear      -0.035661571           1.000000000
```

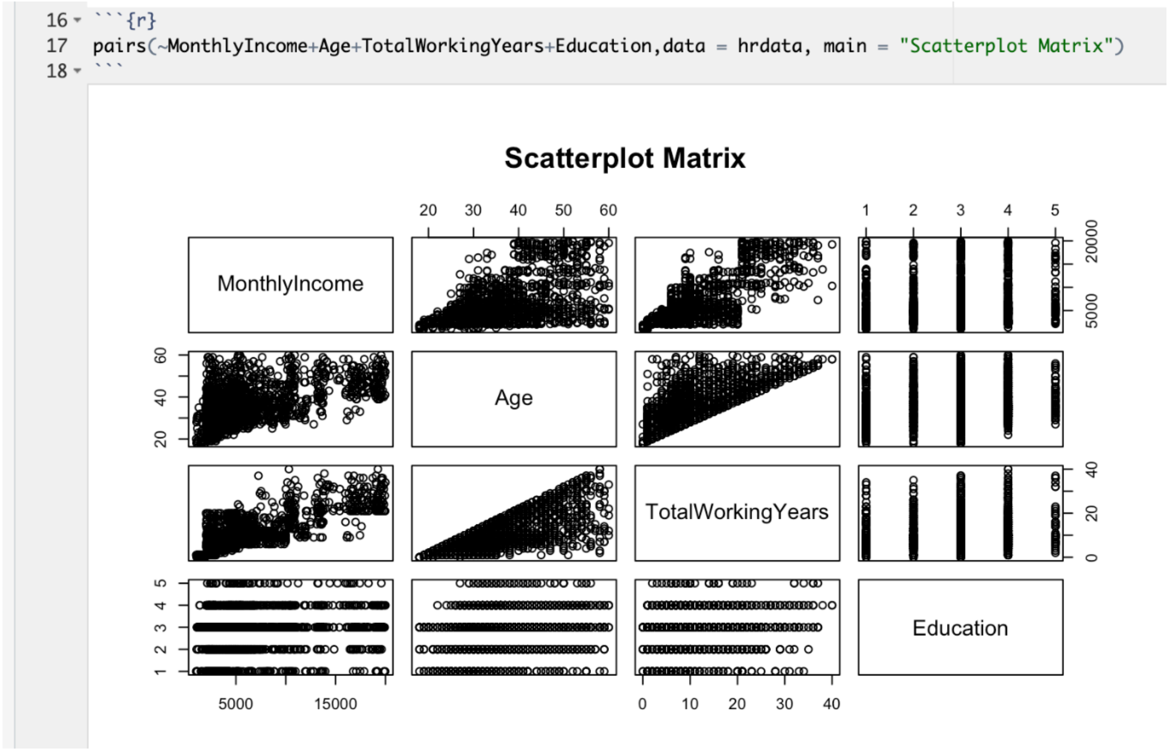
Correlation Matrix

The results revealed significant correlations between:

- Age and Total Working Years
- Age and Monthly Income
- Age and Number of Companies Worked

Visualizing Relationships with Scatter Plots

To visualize these relationships, I created scatter plots using the pairs function in R.

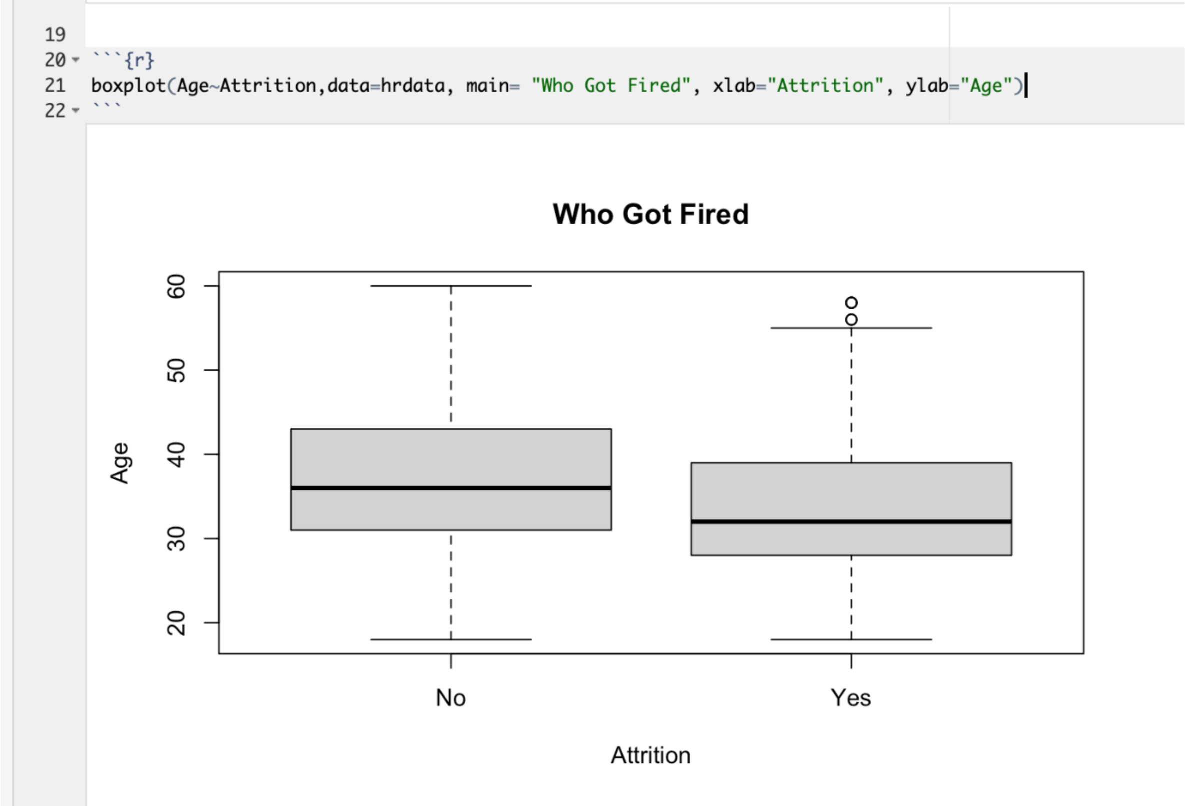


~~~r
```{r}
pairs(~MonthlyIncome+Age+TotalWorkingYears+Education,data = hrdata, main = "Scatterplot Matrix")
~~~
```

Demographic scatterplot

Here are some noteworthy observations from the scatterplot matrix:

- There is a positive correlation between age and total working years, indicating that experience tends to increase with age.
- Monthly income tends to rise with age, although there are a few individuals in their 30s who earn more than those in the subsequent age groups.
```

- The correlation between education and monthly income is weakly positive. This suggests that higher education doesn't always lead to higher income, although some highly educated employees do earn more.

## Hypothesis Analysis: Age and Attrition

To address claims of ageism, I performed hypothesis testing using Welch Two Sample t-tests. The boxplot of age distribution for employees who stayed versus those who left suggested that younger employees were leaving at a higher rate.

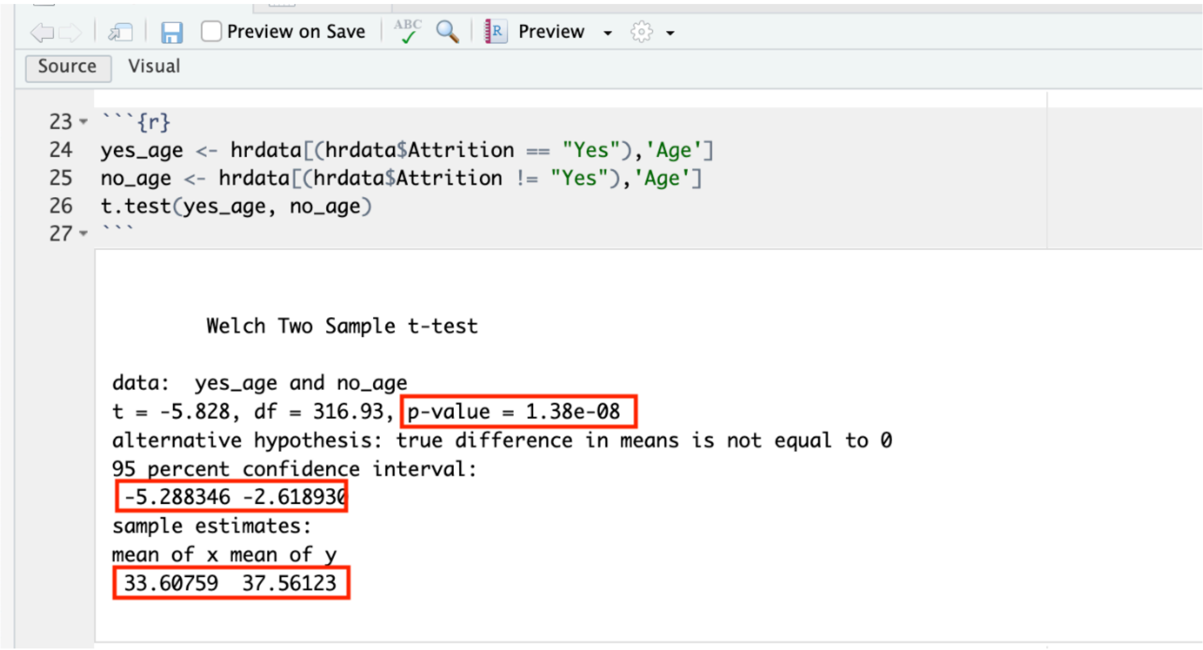


~~~
```{r}
boxplot(Age~Attrition,data=hrdata, main= "Who Got Fired", xlab="Attrition", ylab="Age")
```
~~~

Employee Retention vs Age - Whisker Plots

"Yes" indicates employees who were laid off, while "No" represents current employees. The median age is shown by the line in the middle of the box. Both groups appear quite similar, but the average age of employees who stayed is 38 years, compared to 30 years for those who were laid off.

To confirm our hypothesis, we will perform a two-sample t-test to compare the ages of both groups and determine the p-value.



~~~
```{r}
yes_age <- hrdata[(hrdata$Attrition == "Yes"),'Age']
no_age <- hrdata[(hrdata$Attrition != "Yes"),'Age']
t.test(yes_age, no_age)
```
~~~

```

        Welch Two Sample t-test

data:  yes_age and no_age
t = -5.828, df = 316.93, p-value = 1.38e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.288346 -2.618930
sample estimates:
mean of x mean of y 
 33.60759  37.56123 
```

The t-test results confirmed this observation with a p-value less than 0.05, indicating a significant difference between the two age groups. This suggests that there is no evidential bias, and the employees who left the company are indeed younger than those who stayed.

Key Takeaways

This project highlighted key factors influencing employee attrition at IBM, such as age and working years. Younger employees were more likely to leave, and there was a strong correlation between age and monthly income. By understanding these patterns, IBM can develop targeted strategies to improve employee retention.

If you found these insights valuable, consider how similar analysis can be applied in your organization. Feel free to reach out with questions or for collaboration on similar projects!