Employee Attrition Analysis with R (IBM Dataset)

In the ever-changing business world, understanding and reducing employee attrition is crucial for the success of any organization. Employee turnover not only affects workforce stability but also has far-reaching implications for productivity and performance. In this article, we delve into the world of human resources data analytics, leveraging the power of R, a statistical programming language, to conduct a comprehensive analysis of employee attrition and its impact on performance. By dissecting HR data and employees performance with statistical techniques, we aim to uncover valuable insights into the factors driving attrition and how it relates to workforce performance.

Take this data-driven journey as we explore the patterns, correlations, and actionable strategies that can empower organizations to retain top talent and optimize their performance.

The goal of this analysis is to model employee attrition and determine the most dominant contributing factors that govern this turnover.

About Me

Data Analyst: Shafic Sebanenya
Email: sebanenyashafic2@gmail.com
LinkedIn: Shafic Sebanenya
Contact: +971-56389 5861/ +971-5250 49127

Key Findings of the project:

- There's a strong positive correlation between total years working and monthly income and between age and total years working.
- We found a correlation of 50% between age and monthly income, and this correlation can be attributed to various factors.
- Longevity is not a significant factor in employees' attrition. The length of time an employee stays with the company doesn't play a significant role in their decision to leave.
- There is a strong correlation between education and daily rate, especially in research & development departments.
- Younger workers were consequently more vulnerable to layoffs and were at a higher risk of experiencing attrition by HR as compared to their older counterparts.
- The combined factors of age and the total number of working years account for nearly 60% of the variability in total monthly income.

Let's start with the data:

In this project I act as a People Data Analyst intern for IBM in the Human Resources department. There's been lots of people leaving the company.

The employer has asked me to determine the reasons behind the sudden layoffs of so many employees. A former employee has claimed that ageism played a part in recent layoffs, and my employer wants me to investigate these claims. I've also been asked to look over the data to see if I have any insights I could provide.

The data used for this analysis is an augmented version created by real IBM data scientists, but isn't exactly 100% real data. You can find it here: Dataset. This data set is a mixed metrics of employee information and has

various features. The data set is in CSV format which has 1470 rows and 35 columns. Each row represents an employee. This dataset includes important attributes like:

- Age
- Attrition
- Daily Rate
- Education
- Employee Number
- Gender
- Hourly Rate
- Job Role
- Monthly Income
- Total working years
- Years at company
- Years in current role

The dataset helps in the analytical stage to examine the metrics report to find trends and patterns that could affect a company. Various analytical techniques are employed based on the desired result. Descriptive analytics, prescriptive analytics, and predictive analytics are a few of them. The only goal of descriptive analytics is to comprehend previous data and identify areas for improvement. The goal of predictive analytics is to foresee future dangers or opportunities by analyzing previous data using statistical models.

There are two types of data used in this dataset, Numerical Discrete Data and Text Categorical(Nominal and Ordinal) Data.

Exploration:

For analysis and making visualizations in this project, I decided to use R. First of all I downloaded R and RStudio. I've uploaded this data in R Studio, which is an IDE, Integrated Development Environment of R. I then created a new R Notebook and imported data from the CSV file into RStudio as a data frame called hr_df.
In this analysis we will explore the following,

- Read the data into R
- Evaluate relationships (Correlations)
- Create scatterplots
- Create boxplots
- Hypothesis Testing
- Linear Regression

To see total job roles and employees for these roles, I created a pivot table in Excel. It indicates there are total 9 types of Job titles employees are serving and total number of employees in IBM are 1470.

Row Labels	Count of JobRole
Healthcare Representative	131
Human Resources	52
Laboratory Technician	259
Manager	102
Manufacturing Director	145
Research Director	80
Research Scientist	292
Sales Executive	326
Sales Representative	83
Grand Total	**1470**

A pivot table showing all job titles and total number of employees in these roles

Business Questions:

- What is the extent of correlation between key demographic factors, and how do they collectively impact the company's operations and performance?
- How does the relationship between age and total working years impact the monthly income?
- Do ageism play an important role in employees' layoffs?
- Does higher education level affect employees' daily rate in some specific departments?
- To what extent does the duration of an employee's tenure with the company influence their likelihood of leaving, and how does this finding impact our attrition management strategies?
- How can we predict the monthly income based upon age and total working years?

Analysis:

1- For the first business question, the company wants to get an overview of how some of the most important demographics correlate. I filtered the data so I could pass in all the rows, but just the columns of interest. I created a new vector (hr_df_corr) using the select() function that excludes non -numeric variables using a with following columns to find correlation between them, "Age", "DailyRate", "DistanceFromHome", "Education", "HourlyRate", "MonthlyIncome", "MonthlyRate", "NumCompaniesWorked", "TotalWorkingYears", "TrainingTimesLastYear". I used the cor function below,

```
hr_df <- rename_with(HR.Employee.Attrition,tolower)
View(hr_df)
hr_df_corr <- select(hr_df,"age","dailyrate","distancefromhome","education","hourlyrate"
View(hr_df_corr)
# Is there a correlation between important demographics? If so,how strong is it.
install.packages("corrplot")
library(corrplot)
cor(hr_df_corr)
```

```
                          age    dailyrate distancefromhome   education
age                1.00000000  0.010660943     -0.001686120  0.20803373
dailyrate          0.01066094  1.000000000     -0.004985337 -0.01680643
distancefromhome  -0.00168612 -0.004985337      1.000000000  0.02104183
education          0.20803373 -0.016806433      0.021041826  1.00000000
hourlyrate         0.02428654  0.023381422      0.031130586  0.01677483
monthlyincome      0.49785457  0.007707059     -0.017014445  0.09496068
monthlyrate        0.02805117 -0.032181602      0.027472864 -0.02608420
numcompaniesworked 0.29963476  0.038153434     -0.029250804  0.12631656
totalworkingyears  0.68038054  0.014514739      0.004628426  0.14827970
trainingtimeslastyear -0.01962082  0.002452543     -0.036942234 -0.02510024
                    hourlyrate monthlyincome  monthlyrate numcompaniesworked
age                0.024286543   0.497854567  0.028051167         0.29963476
dailyrate          0.023381422   0.007707059 -0.032181602         0.03815343
distancefromhome   0.031130586  -0.017014445  0.027472864        -0.02925080
education          0.016774829   0.094960677 -0.026084197         0.12631656
hourlyrate         1.000000000  -0.015794304 -0.015296750         0.02215688
```

In this correlation matrix, there is a strong positive correlation between Monthly Income, Age and Total Working Years.

- Age & Monthly Income 49%
- Age & Total Working Years 68%
- Age & Education 28%
- Monthly Income & Total Working Years 77%

We can find in highlighted columns, there is around 50% positive correlation between Age and Monthly Income. As the age increases, it results in an increase in the monthly income of employees, as they gain more experience with time that increases their pay rate. But this is not always the case, sometimes an increase in age does not bring an increase in monthly income. There are several factors that affect this increment like education, skills and experience in a specific field.

Secondly we see 68% correlation between Total Working Years & Age, this relationship seems relatively obvious and can probably be predicted already.

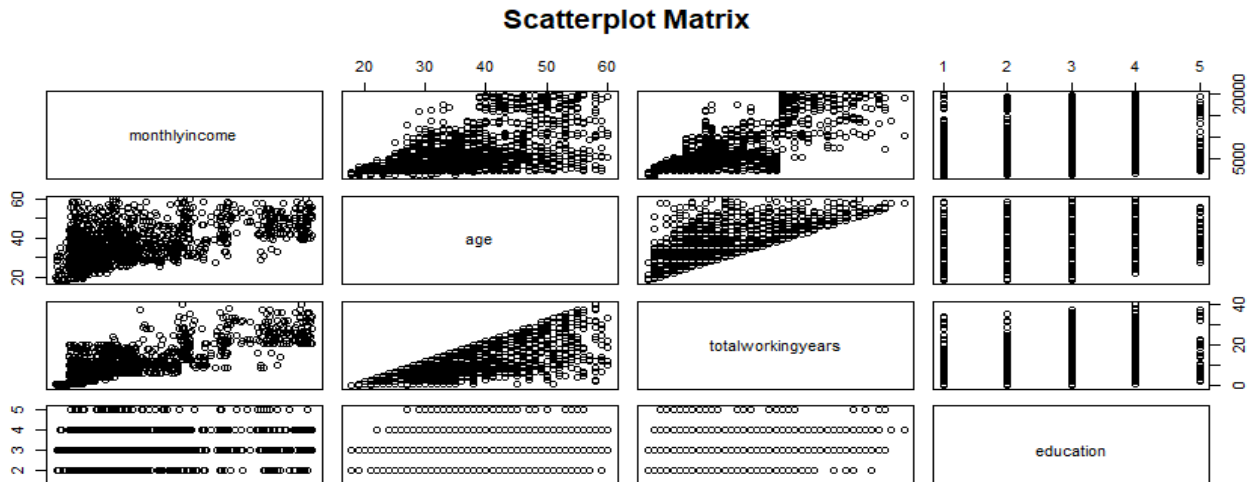
We could find a correlation of 28% between Age and Education, which is not a very strong correlation. That means that with an increase in age all employees do not pursue more education. While some employees continued to get more education with their employment.

We could find a strong correlation of 77% between Monthly Income and Total Working Years. As the experience of employees increases each working year we can see that it adds more value in employees' presence in the company and results in higher income levels.

2- How can we learn more about these correlations? Let's make some scatter plots to further analyze this associated data. To determine the link between MonthlyIncome, Age, TotalWorkingYears, and Education, we will generate scatter plots as Pair Plots in R. As can be seen in the scatter pair plot below, the age-education correlation found in the correlation example above does not translate well. I used pairs() and added 4 columns to see the relationship between them.

```
9  #Scatter pair plots showing relationship between monthly income,age
10 #totalworkingyears and education
11 pairs(~monthlyincome +age + totalworkingyears+education,data=hr_df_corr,
12       main="Scatterplot Matrix")
13
```

Pairplots R code



There is a strong relationship between total working years and Monthly Income and between education and monthly Income. Age and total working years strongly correlate, which is pretty obvious . Age and monthly income show fewer employees have higher monthly income as they get old.

The next business question is to find out if ageism is a factor in layoffs. A disgruntled former employee is suing the business on the grounds that ageism contributed to recent layoffs. They contend that older employees were fired more frequently than younger ones. My boss wants me to use the force of numbers to disprove this notion and stop this prospective litigation. I made the decision to do my research using visualization and hypothesis testing.

I analyzed and created a box plot to compare the age ranges of those employees who got laid off. As seen by the lines in the middle of the plots below, a greater age appears to be marginally more favorable. The average age of those who were kept on staff was a little older than that of those who were let go.

```
13  # A box whisker plot to see if age is a factor in employee attrition.
14  boxplot(age~ attrition,data= hr_df, main= "Who got Fired", xlab="attrition
15          ylab="age")
```

Box whisker plot R code

Currently, they appear to be quite close at first glance, particularly the median numbers. However, as we can see, the "Yes" attrition median number is actually a little lower than the "No", indicating that the average age of the departing employee may be lower than the average age of the retained employees. How much though? And does that matter? There are surely many other factors affecting the attrition rate.

So the next business question is whether the outcomes of this one layoff are statistically significant. Since this is not a circumstance like a medical trial, I will use a statistically significant 95% confidence level since 95% is normally recognized as certain unless it's a life or death matter to assess the data.

We can't tell from a box plot, so let's test the hypothesis. Another statistical technique for reaching judgments or inferences about a population from a sample of data is hypothesis testing. Based on statistical evidence, hypothesis testing enables us to take wise judgments and reach meaningful conclusions. It is commonly utilized to obtain understanding and confirm population assumptions in research, experiments, quality control, and many other fields.

We have created the variables yes_age for employees who were fired and no_age for employees who were able to keep their employment in the following query. We'll be able to obtain the findings using the t-test function.

A p-value represents the likelihood that more severe data would be seen if the null hypothesis (no difference between groups) were true. By informing us of the strength of the evidence opposing the null hypothesis, it helps us make decisions during the hypothesis testing process. A higher p-value indicates less convincing evidence, whereas a lower p-value indicates more convincing evidence that the null hypothesis is untrue.

The Welch two-sample t-test will be used to compare the mean ages of the two groups. By defining variables called yes_age and no_age, I must first determine the mean ages of the employees who were kept and the mean ages of the employees who were let go during this attrition process.

To do this in R, I created a new variable called yes_age that is the Age column but only the rows that have attrition as "Yes".

Then create another variable called no_age that is the Age column, but only the rows that have "No" in attrition. and use these two variables as arguments in the t-test.

```r
# To compare average ages and to calculate p-values, I created two variables  called
#yes_age & no_age

yes_age <- hr_df[(hr_df$attrition == "Yes"),'age']
no_age<- hr_df[(hr_df$attrition != "Yes"),'age']

# Running the t-test
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
```

Two sample t-test showing the p-value

This tells us that the mean age of the group that got laid off is 33.6 and the mean age of the retained workers is 37.56. The p-value of is very small. **Because p-value is less than 0.05**, there is a statistically significant difference between the two samples, but not what that angry previous employee claimed. Those who left were younger than those who stayed! We can see that in the mean comparison at the bottom. **We can reject the null hypothesis** that the true difference in means is zero and accept that we have convincing statistical evidence that there is an age difference between retained and laid-off workers. Here x is the first array we passed in and y is the second array we passed in. That is also confirmed in the confidence interval, since both those numbers listed below are negative, we know that the first array is smaller than the second, with confidence.

Because the business now has proof to disprove the accusation raised in the potential litigation, the boss is pleased with our results. To find out why younger workers were let go more frequently, they might wish to look into the matter more.

Perhaps it has to do with their lack of experience and the need to retain the most experienced workers.

5- Is Longevity a Factor in the Layoffs?
Another disgruntled employee states that layoffs were just based on the EmployeeNumber, and new employees were let go more than employees with greater longevity in the company.

I repeated my analysis, switching Age for EmployeeNumber and created a boxplot and ran a t-test.

I used this command to generate a boxplot:

```
29 # A Box whisker plot to see if employee number is a factor in employee attrition, and
30 # new employees were let go more than old employees.
31 boxplot(age~ attrition, data=hr_df, main= "Who got fired", xlab="attrition",ylab="employmentnumber")
```

I obtained the boxplot, which revealed that there was hardly any difference between the two groups' medians. Only a few modest differences between new and seasoned personnel are found, according to the analysis. I'd say it seems unlikely that this variable affects worker attrition.
Let's see the p-value for this two sample t-test.

```
34 #created two variables  called yes_employeenumber and no_employeenu
35 #to compute the p-values
36 yes_employeenumber <- hr_df[(hr_df$attrition == "Yes"),'employeenum
37 no_employeenumber<- hr_df[(hr_df$attrition != "Yes"),'employeenumbe
38
39 #Running sample t-test
```

```
        Welch Two Sample t-test

data:  yes_employeenumber and no_employeenumber
t = -0.41725, df = 342.33, p-value = 0.6768
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -98.91087  64.29061
sample estimates:
mean of x mean of y
 1010.346  1027.656
```

We may conclude that longevity did not affect layoffs given that the confidence interval comprises zero as a likely value for the genuine difference in means between the two groups and that there is a 67.7% possibility that a difference in means this large could happen by chance alone.

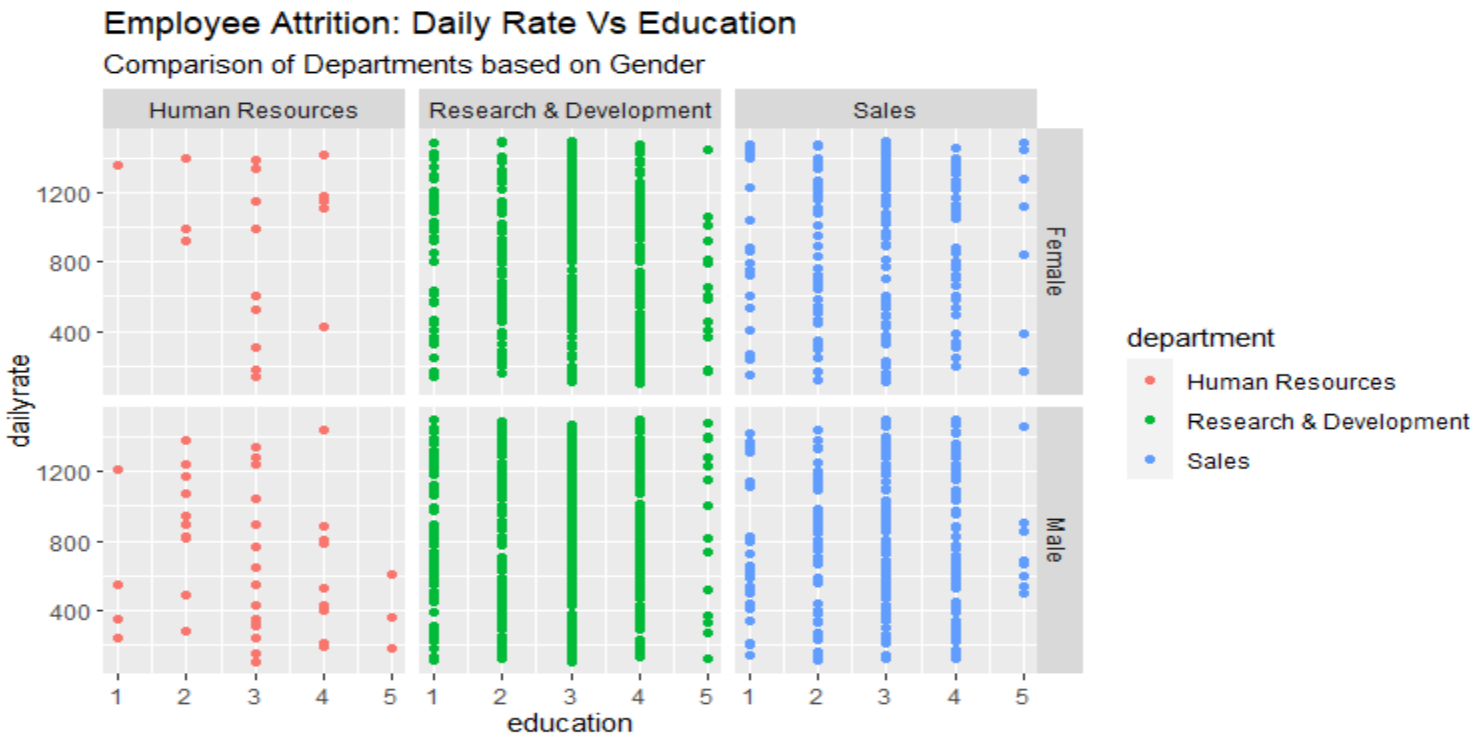
6- Our next business question is to find if there is correlation between education and daily rate. To know about this I installed and loaded ggplot2 package using following functions

```
# Installing and loading ggplot2 package
install.packages("ggplot2")
library(ggplot2)
```

Using geom_point() and facet_grid() I created scatter plots. facet_grid function can show and quickly organize complex data and make it easy to spot relationships. I've assigned columns Education on x-axis and Daily Rate on y-axis. By adding color to Departments, the color legend shows us 3 departments with Human resources in red color, research and development in green and sales in blue color. I've also added a 4th variable, Gender in facet_grid to see if there is any specific pattern found in men employees vs women employees.

```
#Scatter plots to find relationships between education and dailyrates
#and to analyze if gender or dept is a factor in employee attrition
ggplot(data=hr_df)+
  geom_point(mapping=aes(x=education, y=dailyrate, colour= department
                         )) +facet_grid(gender~department)+
  labs(title= "Employee Attrition: Daily Rate Vs Education",subtitle =
         "Comparison of Departments based on Gender")
```

Scatter plot R code using ggplot2 package



Scatter plot showing relationships between Daily Rate Vs Education

The daily rate and education level are strongly correlated. We can see that as education levels rise, so does the impact on employers' daily rates. In the research and development department, I can observe that highly educated personnel have higher daily rates, particularly for men. Based on their deduction and daily rate, there is no prominent gender discrimination in the attrition rate of employees found. IBM has rigorous policies regarding racial and gender discrimination in the workplace.

7- How we can predict monthly income based upon age and total working years. We can perform multivariate regression to more accurately predict someone's monthly income with both age and total years working as explanatory variables.

```
54  #A linear regression model that predicts the Monthly Income based upon Age
55  model1 = lm(monthlyincome ~ age,data= hr_df)
56  summary(model1)
```

Linear Regression R code

```
Residuals:
    Min      1Q  Median      3Q     Max
-9990.1 -2592.7  -677.9  1810.5 12540.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2970.67     443.70  -6.695 3.06e-11 ***
age           256.57      11.67  21.995  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4084 on 1468 degrees of freedom
Multiple R-squared:  0.2479,    Adjusted R-squared:  0.2473
F-statistic: 483.8 on 1 and 1468 DF,  p-value: < 2.2e-16
```

Linear Regression R code

The multiple R-squared value above is 0.2479. That means that 25% of monthly income can be explained in relation to age, which is note a very strong relationship. The idea that as we get older, our monthly salary will also increase is not really realistic. The p-value is 2.2e16, which is equivalent to zero (e is 10 to the power). We can declare with 95% certainty that this model is statistically significant because p is less than 0.05. Here we can add another variable, TotalWorkingYears, and see how it correlates with age and if there are any changes in the results.

```
58 #A linear regression model that predicts the monthly income based upon age and total
59 # working years.
60 model2 = lm(monthlyincome~ age + totalworkingyears, data= hr_df)
61 summary(model2)
```

Linear Regression R code

```
Call:
lm(formula = monthlyincome ~ age + totalworkingyears, data = hr_df)

Residuals:
     Min       1Q   Median       3Q      Max
-11310.8  -1690.8    -91.4   1428.3  11461.5

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        1978.08     352.36   5.614 2.36e-08 ***
age                 -26.87      11.63  -2.311    0.021 *
totalworkingyears   489.13      13.65  35.824  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2984 on 1467 degrees of freedom
Multiple R-squared:  0.5988,    Adjusted R-squared:  0.5983
F-statistic:  1095 on 2 and 1467 DF,  p-value: < 2.2e-16
```

The multiple R-squared value above is 0.5988. We can see that about 60% of the age variant can be explained in correlation with total working years. A substantial connection, R-squared of 59%, indicates that the number of working years will increase as an employee becomes older and if they continue to work. The p-value is 2.2e16 once more, or essentially 0. We can say with 95% confidence that this model is statistically significant and an even better predictor of monthly income because p is less than 0.05.

Age and total working years –explain almost 60% of the variation in total monthly income. This makes sense. Some older workers have jobs with lower earning potential and fewer educational and professional requirements. However, 60% of the variation in monthly income can be attributed to age and years of employment.

Insights & Recommendation:

It seems as though IBM doesn't engage in discrimination. Ageism and gender bias don't seem to play a role in decisions about who gets retained after a layoff. Surprisingly, despite the fact that performance evaluations are highly connected with pay raises, neither do performance reviews.

The T-test demonstrated that younger people have moved away in search of better prospects. People who departed on average were 33.6 years old, while those who stayed on average were 37.6 years old. Obtain information to comprehend the objectives and aspirations of each employee, paying particular attention to younger workers. Professional training and an environment that will support their growth will best increase the retention rate of younger employees.

In general, older employees have more experience and have worked longer. Although there are some issues with the relationship, it is more likely that they have moved up the corporate ladder and are now in positions with higher earning potential. There are also other factors in play. Because they changed careers, took time off to start families, further their education, or for other reasons, some older employees had fewer years of employment than others.

To maintain a fair and inclusive work environment, it's essential to continue enforcing IBM's robust policies against racial and gender discrimination. Additionally, consider implementing educational and career development programs that empower employees to further their education and skills, potentially contributing to higher daily rates and career advancement. This can enhance overall job satisfaction and retention while promoting diversity and inclusion within the organization. A copy of the R script is available

Conclusion:

Thank you for reading my project article of Statistical Analysis of HR Data with R. If you have any feedback or questions about my insights and analysis, feel free to contact me. You can reach out to me at

sebanenyashafic2@gmail.com.

LinkedIn : Shafic Sebanenya