**Predicting Employee Retention Using Logistic Regression by Ayesha Imran (SP23-BBD-011)**

**Table of Contents**

1. **Abstract………………………………………………………………………2**
2. **Introduction………………………………………………………………….2**  
   2.1. Overview of Employee Retention Challenges……………………………2  
   2.2. Importance of Predictive Analytics in HR………………………………..2  
   2.3. Objective of the Research………………………………………………….2
3. **Methodology………………………………………………………………….2**  
   3.1. Data Collection and Preprocessing……………………………………….2  
   3.2. Model Development and Training………………………………………..3  
   3.3. Model Evaluation and Optimization………………………………………3
4. **Results…………………………………………………………………………3**  
   4.1. Descriptive Statistics……………………………………………………..3  
   4.2. Model Performance Metrics……………………………………………….3
5. **Discussion……………………………………………………………………..3**
6. **Conclusion…………………………………………………………………….4**
7. **Additional Information………………………………………………………4**  
   7.1. Confusion Matrix………………………………………………………….4  
   7.2. Accuracy Scores…………………………………………………………..4  
   7.3. Graphical Analysis…………………………………………………………5-15
8. **References……………………………………………………………………..16**

**Abstract**

This study utilizes logistic regression to predict employee retention based on HR-related features such as salary, job satisfaction, and tenure. The data underwent preprocessing and feature selection to prepare it for model training and evaluation. Mean Squared Error (MSE)of this model is 0.5 and Root Mean Squared Error (RMSE) is 0. 7071067811865476.The model achieved an accuracy of 83.33%, demonstrating good performance in identifying employees likely to remain but requiring improvements in predicting those likely to leave. The confusion matrix highlighted these strengths and areas for refinement. Future efforts could enhance the model's accuracy and applicability by addressing data imbalances and expanding feature sets.

**Introduction**

**Overview of Employee Retention Challenges**

Employee retention is a critical issue for organizations, directly impacting operational efficiency and profitability. High turnover rates increase recruitment costs, disrupt workflow, and adversely affect team morale. Identifying and mitigating factors that lead to employee attrition are paramount for HR departments.

**Importance of Predictive Analysis in HR**

Incorporating predictive analytics allows HR professionals to proactively address employee retention challenges. By leveraging machine learning models, organizations can uncover patterns in employee behavior, predict turnover risks, and implement targeted strategies to retain top talent. Logistic regression, a widely used algorithm, offers simplicity and interpretability, making it an effective tool for analyzing workforce data.

**Objective of the Research**

This study aims to build and evaluate a logistic regression model to predict employee retention. By analyzing HR datasets, the research seeks to identify key factors influencing attrition and provide actionable insights for workforce management. The ultimate goal is to enhance decision-making processes within HR departments through data-driven methodologies.

**Methodology**

**3.1 Data Collection and Preprocessing**

* **Data Source**: Employee dataset containing demographic, performance, and tenure details.
* **Data Cleaning**: Handled missing values and normalized numerical variables.
* **Feature Engineering**: Selected features with the highest impact on employee retention, such as salary, job satisfaction, and tenure.

**3.2 Model Development and Training**

* **Algorithm**: Logistic Regression, chosen for its simplicity and interpretability.
* **Training**: The dataset was split into training and testing subsets, with 70% allocated for training.

**3.3 Model Evaluation and Optimization**

* **Metrics**: Accuracy, Confusion Matrix, and Classification Report were used to assess model performance.
* **Optimization**: Identified areas where the model underperformed and proposed adjustments for improvement.

**Results**

**4.1 Descriptive Statistics**

* **Dataset Overview**: Employee demographics showed diverse representation across age, gender, and job roles.
* **Key Features**: Salary and job satisfaction emerged as the most significant predictors of retention.
* **Graphs**: We drew and plotted different graphs to check the relation between different variable and attrition.

**4.2 Model Performance Metrics**

* **Accuracy**: 83.33%
* **mean square error of model:**0.5
* **root mean square error of model:**0. 7071067811865476
* **Confusion Matrix**: Highlighted strong performance in predicting employees who would stay but underperformance in predicting those likely to leave.

**Discussion**

The logistic regression model showed adequate performance, with an accuracy score of 83.33%. The confusion matrix indicated room for improvement, particularly in predicting employee attrition.

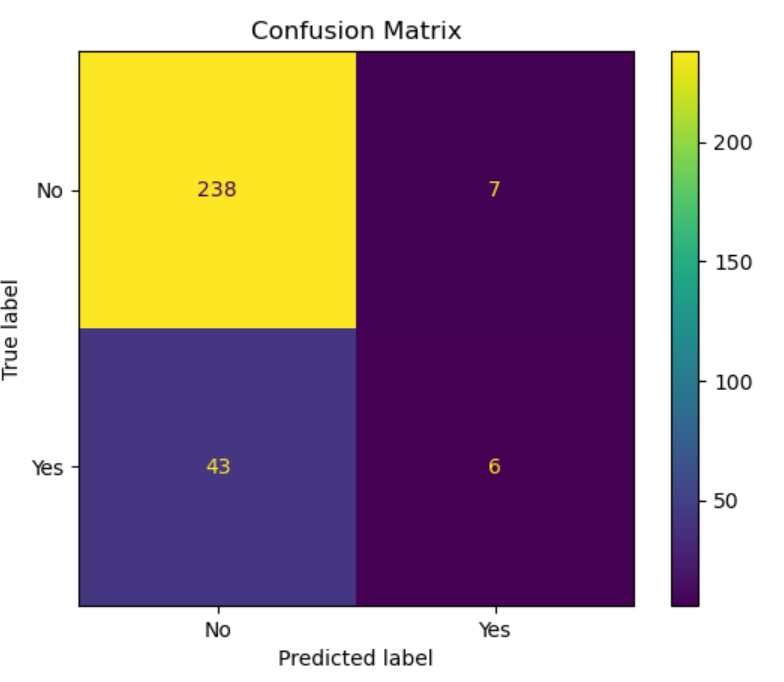
**Conclusion**

This study demonstrates the potential of logistic regression in predicting employee retention. While the model achieved promising results, further refinements could enhance its accuracy and applicability for HR decision-making.

**Additional Information**

**7.1 Confusion Matrix**

The confusion matrix visualized the model’s strengths and weaknesses, emphasizing its accuracy in predicting retained employees.

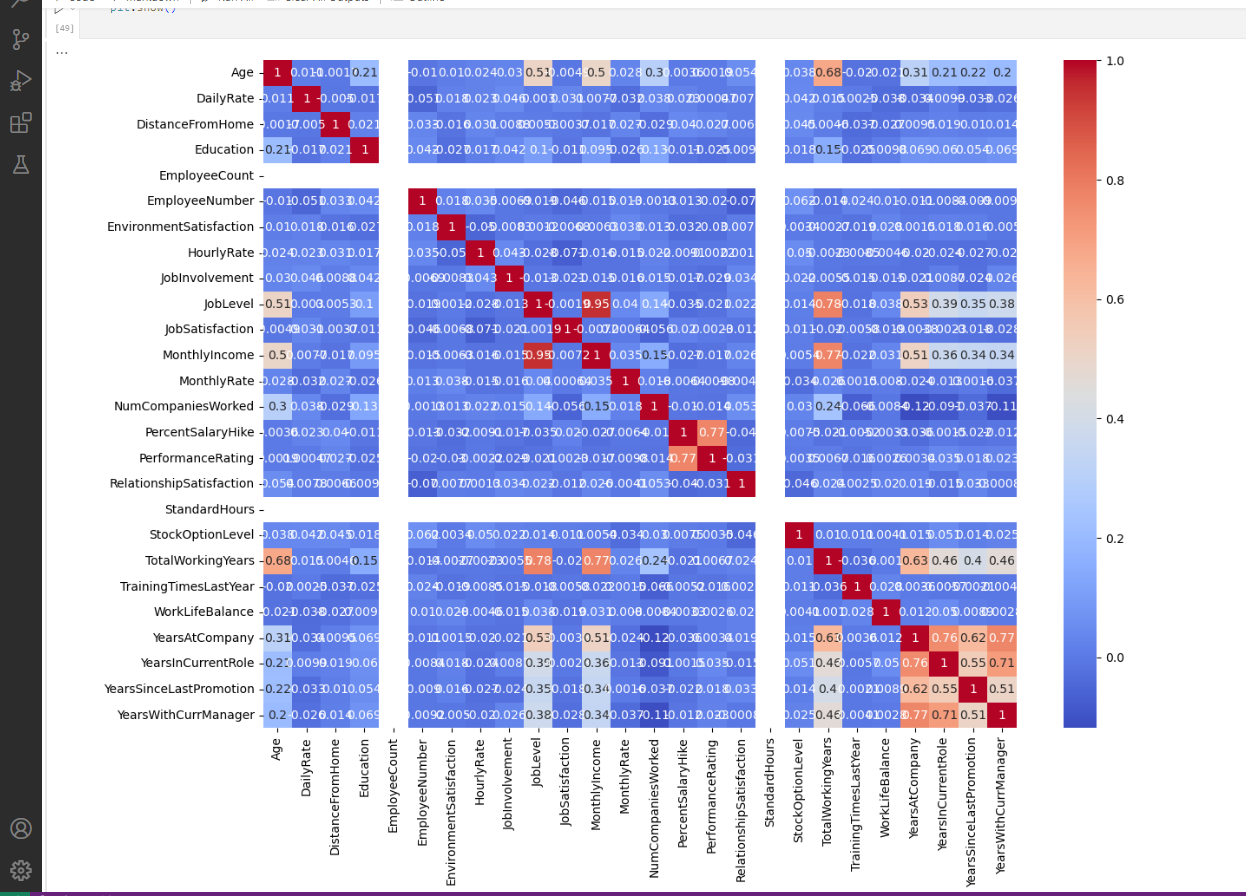


**7.2 Accuracy Scores**

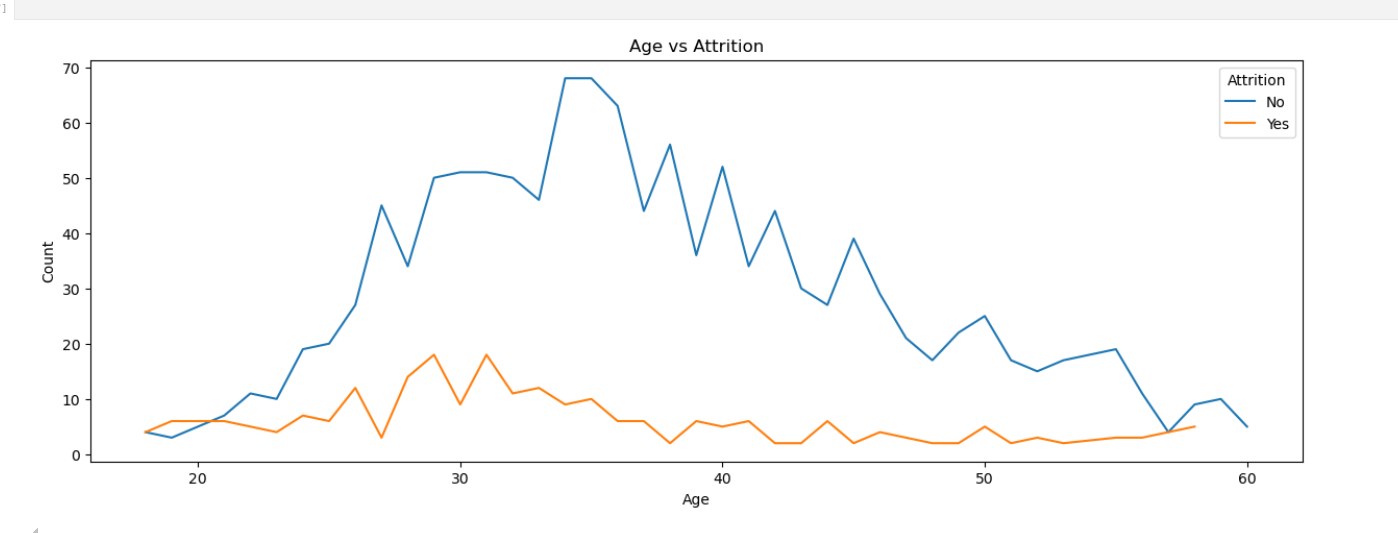
* **Achieved Accuracy**: 83.33%
* **Potential Improvements**: Optimizing model parameters to reduce errors.

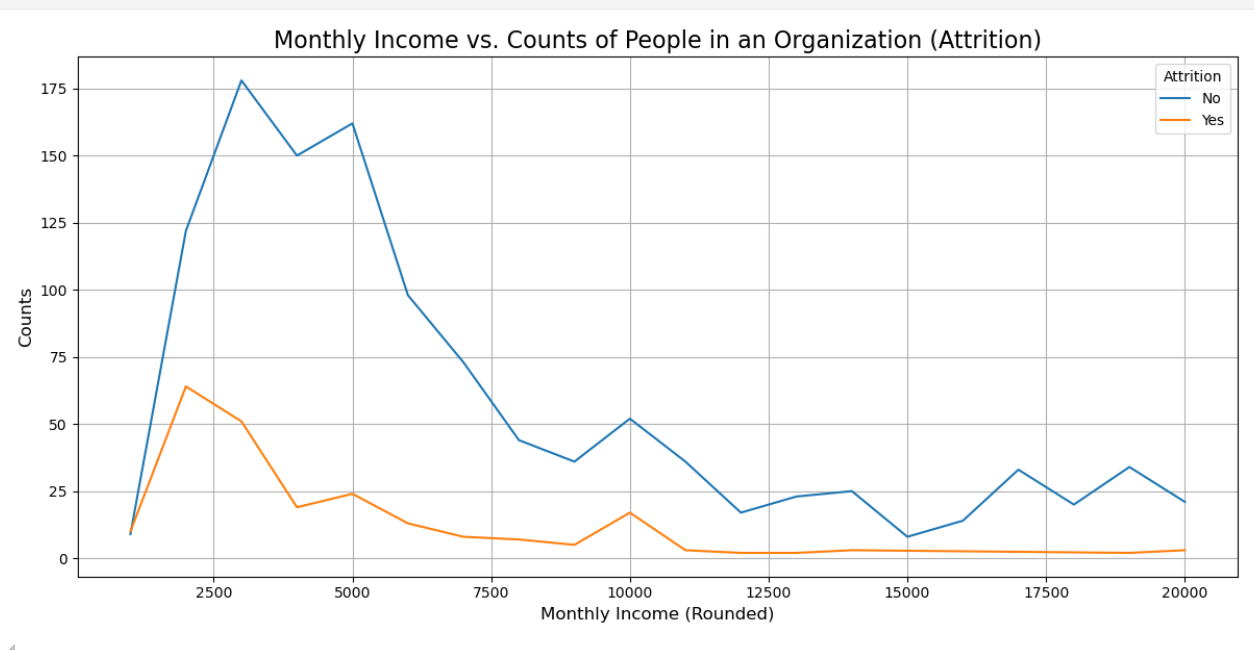
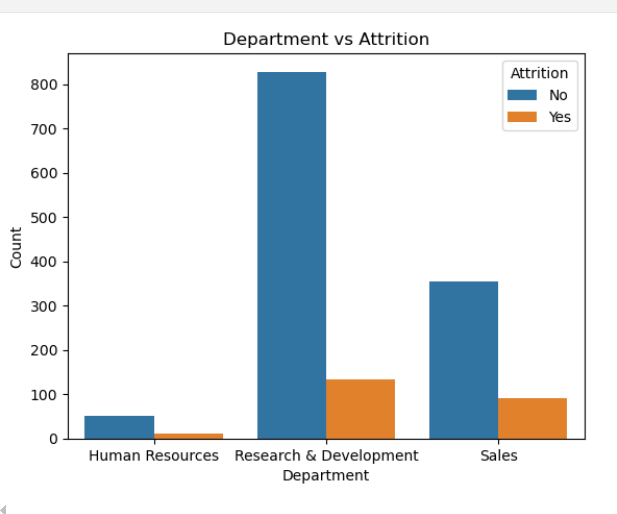
**7.3 Graphical Analysis**

Visuals such as feature importance graphs and confusion matrices provided insights into the model’s performance and the influence of various predictors.

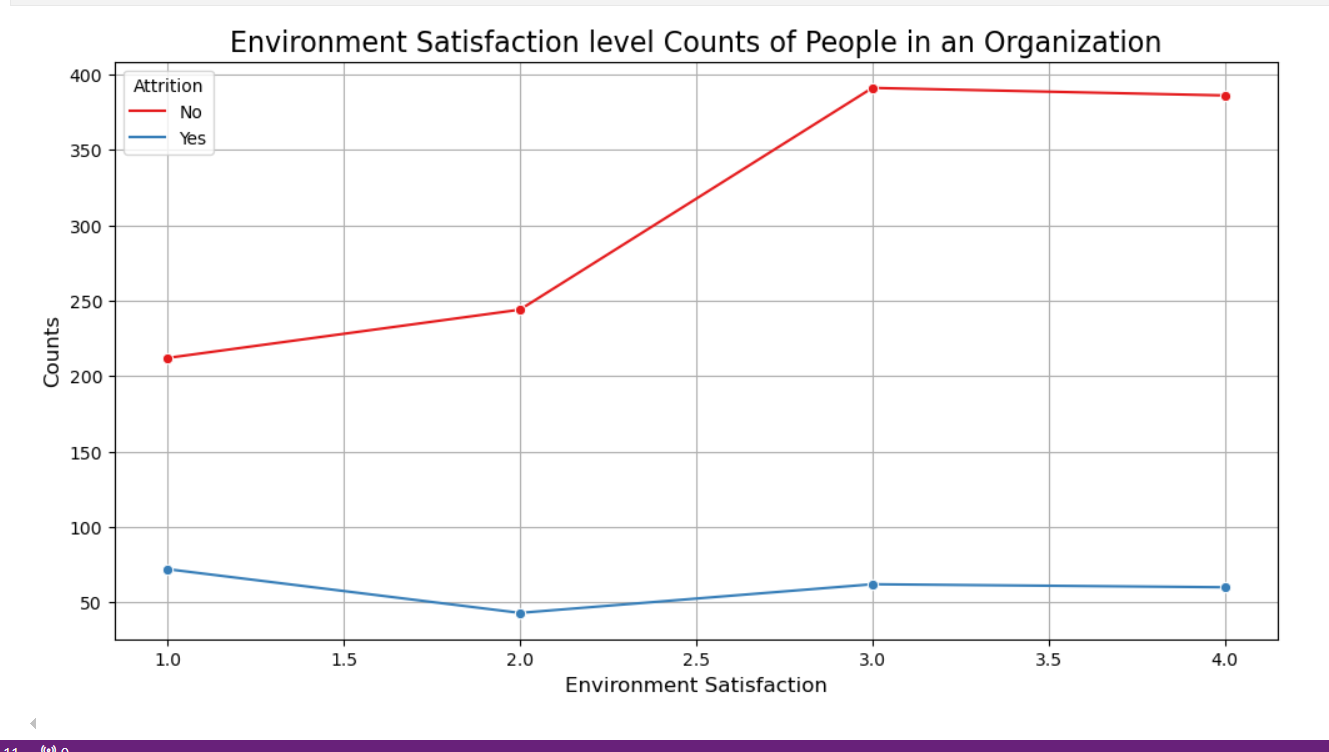


Correlation is a statistical measure that indicates the degree and direction of the relationship between variables. It shows how one variable might change in response to changes in another variable. A correlation value close to **+1** or **-1**is indicating a strong relationship. A value close to **0**is indicating little or no relationship.

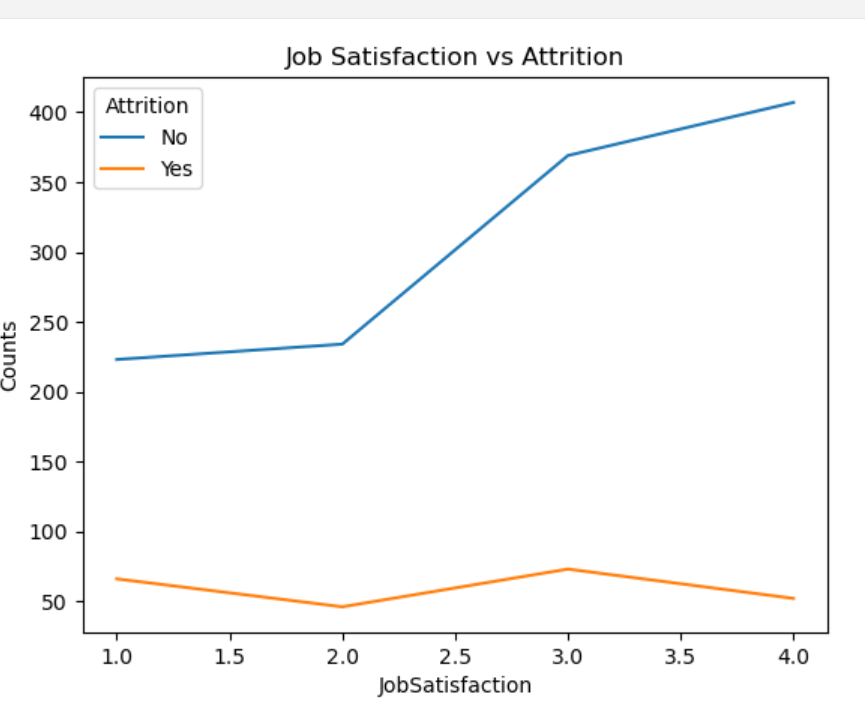
This graph shows that People with age 28 to 32 are mostly going for attrition while at the later ages there is a stability and people are less seen to gone from the office.

Attrition rate is evidently high at very low-income levels- less than 5k month. 

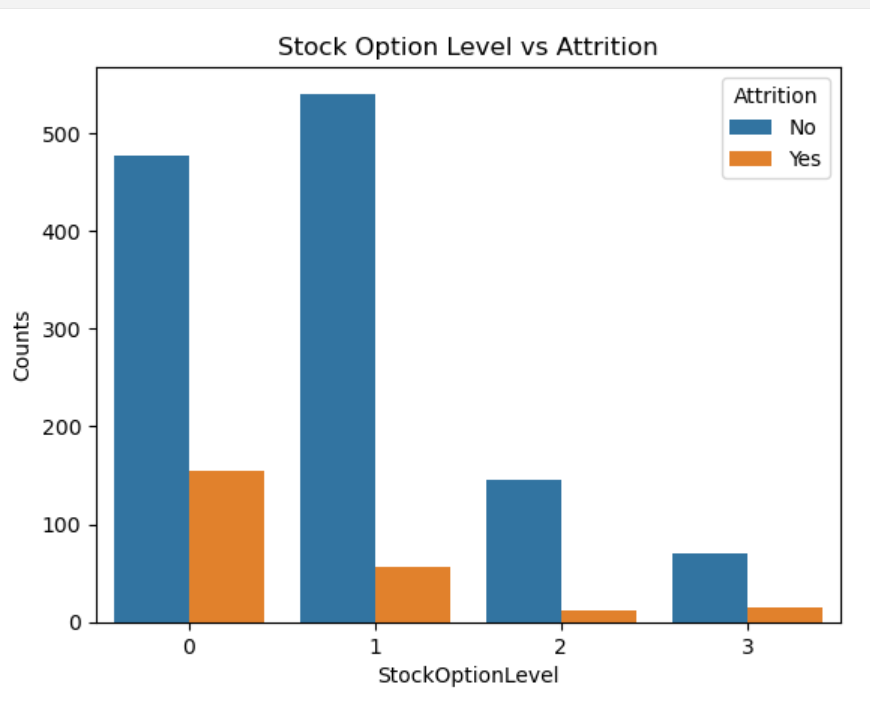
The bar plot shows that the research development department has the highest attrition followed by sales and then at last human resources.



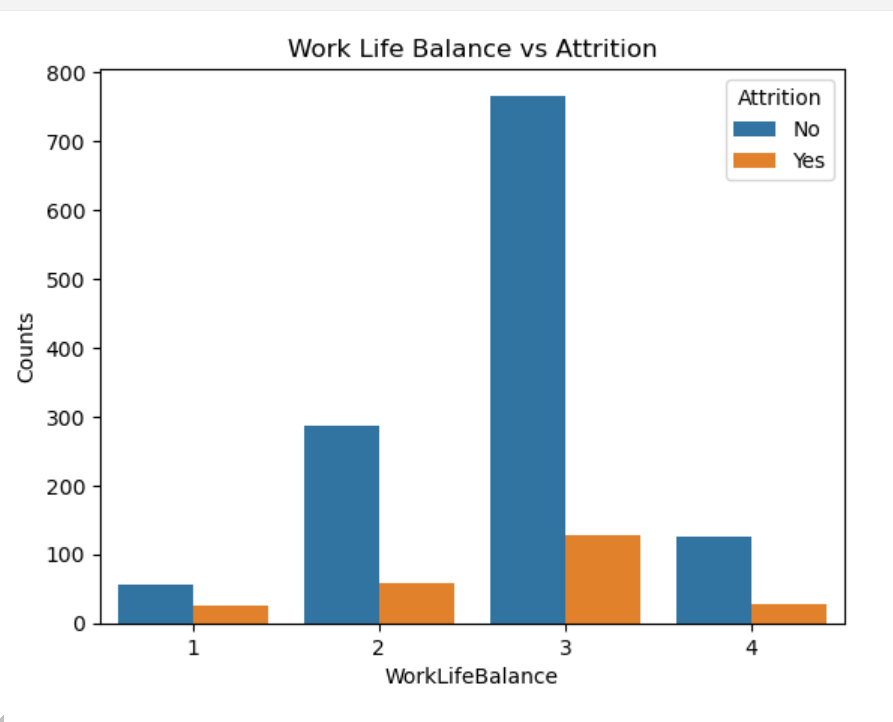
In the satisfaction Level 1-2, the chances of people leaving the organization slightly decreases. This is indicative of the better hopes with which people stay in an organization. However, as we move from 2-3, people tend to move on to get better opportunities and experiences. The attrition rate is almost stagnant for the higher satisfaction levels.



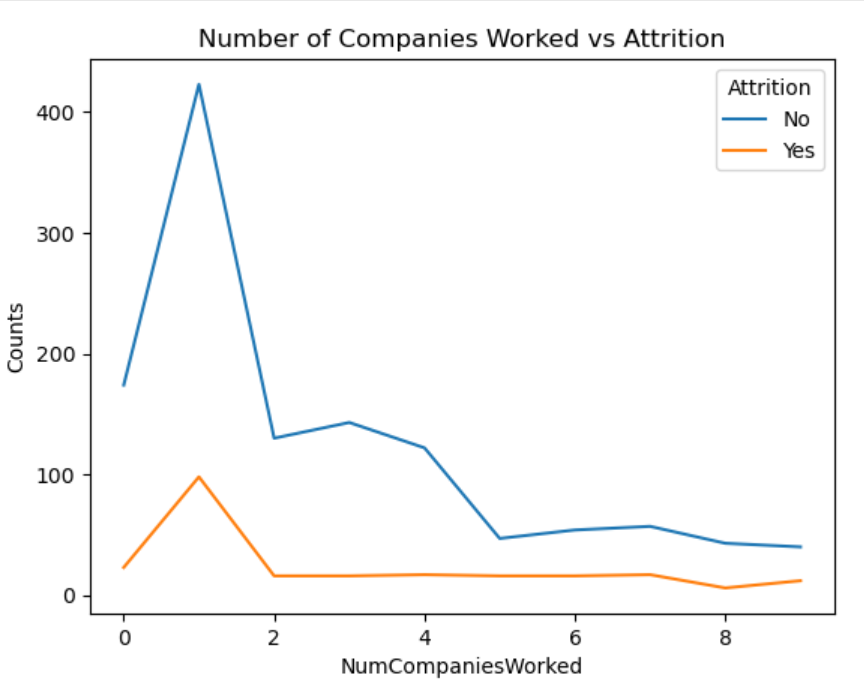
With an increasing job satisfaction, the attrition rates decrease as can be seen in the chart above. Also, from range 1-2 range we can infer (as seen above in Environment Satisfaction), the attrition level falls, but raises from 2-3, where the people tend to choose better opportunities.



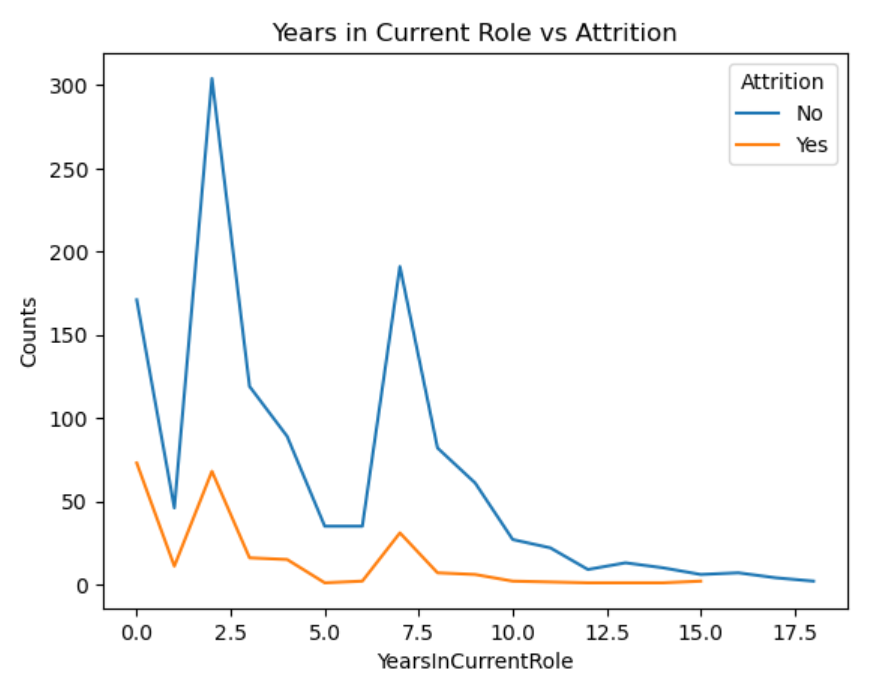
The bar plot shows that the tendency of employees to leave the organization is much more when the stock availing options are limited. Since the stocks constitute to a huge amount of money while staying for a few years, people do not want to lose that opportunity. People with very limited/no stockoptions have a freedom to leave the organization at will.



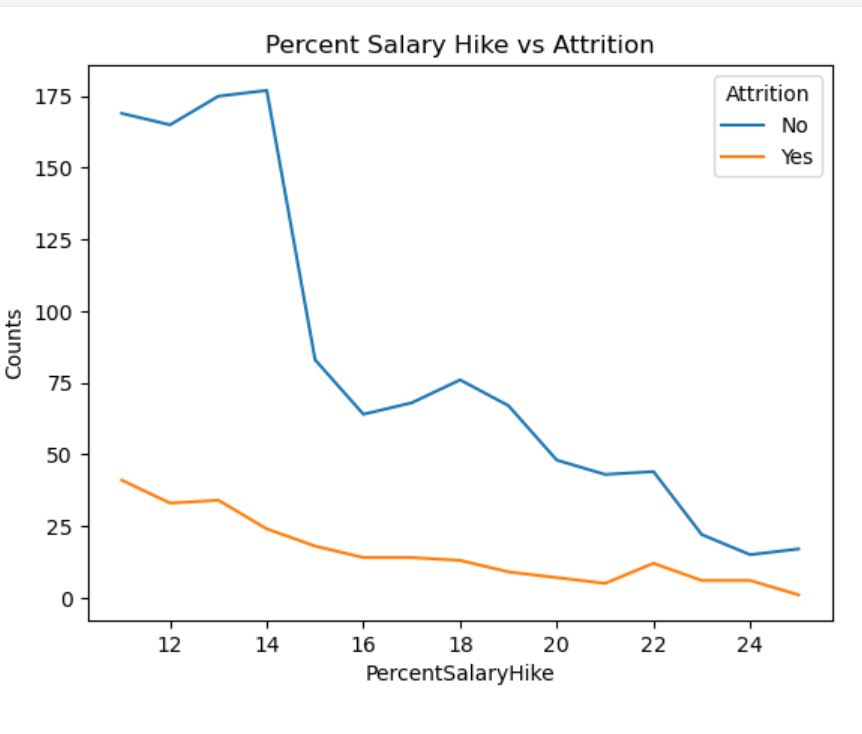
Bar plot shows that the people with poor levels of Work life balance have adjusted themselves to their jobs, but as seen for the above parameters with a better work life score, people are looking towards the better life and want to go for an attrition more. But this trend perishes when the work life balance is really good, and people are satisfied with the work they are doing



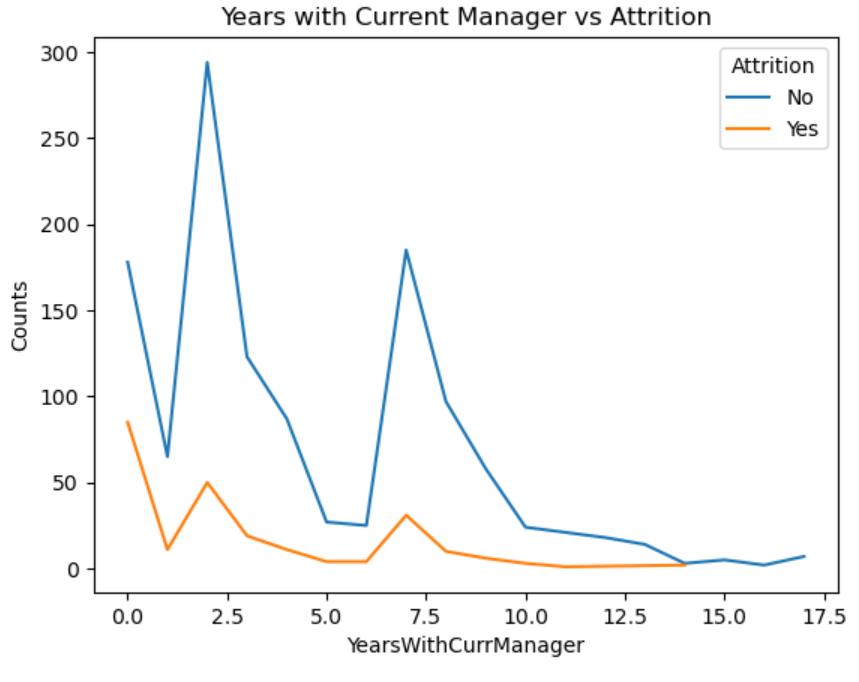
As seen from the chart above, clearly, employees who started their career with the company- or have switched to the company in the initial years of their career, have a higher chance of leaving the organization to a different company. People who have gained much experience- working in multiple companies tend to stay in the company they join.



We have seen people are more prone to leave the organization in the starting years on their role. When people are in the same role for a long period of time, they tend to stay longer for moving in an upward role.



The graph shows that the higher hikes motivate people to work better, and stay in the organization. Hence, we see the chances of an employee leaving the organization where the hike is lower, is much more than a company that gives a good hike.



The graph shows that there are three major spikes. first is as people are working with a new manager, they tend to attrition more quickly the second is after spending 7 8 years they tend to change. they might do this to have some more other perspective or mentorship and after that the attrition rate is much less as compared to initial stages.

**References**

1. Malik, E., Baig, S.A. and Manzoor, U., 2020. Effect of HR practices on employee retention: The role of perceived supervisor support. *Journal of Public Value and Administrative Insight*, *3*(1), pp.1-7..
2. PARK, Y. and Lee, D.G., 2021. Development of a resignation prediction model using HR data. In *Proceedings of the Korean Institute of Information and Commucation Sciences Conference* (pp. 100-103). The Korea Institute of Information and Commucation Engineering.
3. PARK, Y. and Lee, D.G., 2021. Development of a resignation prediction model using HR data. In *Proceedings of the Korean Institute of Information and Commucation Sciences Conference* (pp. 100-103). The Korea Institute of Information and Commucation Engineering.