Employee Attrition in Marvelous Construction

Analysis Report

Problem Overview

Employee attrition has become a significant concern for Marvelous Construction, with many employees resigning from the company. High attrition rates can negatively impact company productivity, morale, and financial stability. The CEO has commissioned a data analysis project to address this alarming situation to understand the underlying reasons behind employee attrition and develop strategies to improve employee retention.

Dataset Description

The dataset provided by Marvelous Construction contains employee details, attendance records, leave information, and salary data. It consists of the following files:

1. **Employee File:**
   Having the information on employee, such as Employee_No, Employee_Code, Name, Title, Year_of_Birth, Gender, Religion_ID, Marital_Status, Designation_ID, Date_Joined, Date_Resigned, Status, Reporting_emp_1, Reporting_emp_2, Employment_Category, Employment_Type, Religion, Designation, etc.

2. **Leaves File:**
   Includes details about employee leaves, such as Employee_No, leave_date, Type (Half day/Full Day), Applied Date, Remarks, apply_type (Annual/Casual).

3. **Salary File:**
   Contains the salary information of employee, including Employee_No, Amount, month, year, and potentially other factors.

4. **Attendance File:**
   Includes attendance data with fields like id, project_code, date, out_date, employee_no, in_time, out_time, Hourly_Time, Shift_Start, Shift_End.

Data Pre-processing

1. **Handling Missing Values:**
   The code calculates the number of missing values in each column using the **isnull().sum()** function. For the 'Marital_Status' column, which had missing values, the code uses a **Decision Tree classifier** to predict the missing values based on other relevant features like 'Gender,' 'Religion,' 'Designation,' and 'Year_of_Birth'. The Decision Tree classifier is trained on non-missing data to make accurate predictions for missing values.

2. **Handling Duplicate Values**
   The preprocessing step identifies and checks for duplicate values in the data frame. While it calculates the total number of duplicate rows, the code doesn't automatically remove them from the dataset. There were no such duplicate rows in the dataset.

3. **Handling Non-Meaningful Values:**
   The code replaces non-meaningful values in the 'Year_of_Birth' column, such as **"'0000'",** with **'N/A'** to indicate unavailable birth years. The code then converts the "Year_of_Birth" column to a numeric format, ignoring any non-numeric entries. It then calculates the mean birth year for married and single employees and fills in missing birth years in their respective categories with rounded mean values. This ensures data consistency and accuracy in analyzing the age distribution by marital status.

4. **Handling Inconsistencies:**
   the code handles "Date_Resigned" and "Inactive_Date" columns, replacing **'0000-00-00'** values with **'\N'** to represent missing dates consistently. It also correctly transfers resignations from "Date_Resigned" to "Inactive_Date" for inactive employees and **vice versa**. These preprocessing steps help ensure that the data is consistent.

5. **Handling Data Integration:**
   The integration of data from multiple files (employee, leaves, salary, and attendance files) is mentioned in the code comments using the common field **'Employee_No'**.

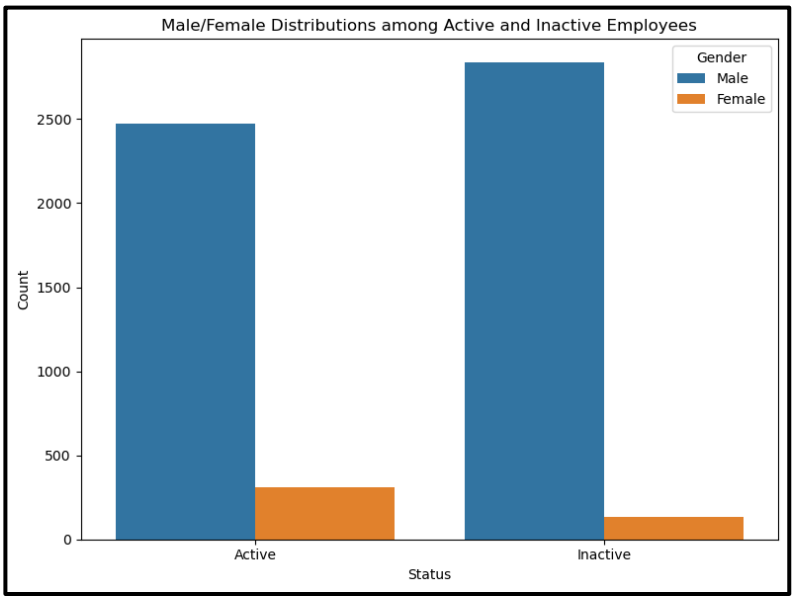
6. **Encoding Categorical Variables:**
   Categorical variables like 'Gender,' 'Religion,' and 'Designation' are **one-hot encoded** to convert them into a suitable format for machine learning models. The OneHotEncoder class from sklearn.preprocessing is used to perform one-hot encoding.

7. **Handling Outliers:**
   Handling outliers, which is an essential step for data preprocessing. Outliers can be detected and managed using various techniques, such as visualizations like box plot, statistical tests, or outlier detection algorithms.

Insights from data analysis

Insight 1: Relationship between gender and retention

Fostering an inclusive and equal workplace requires resolving any differences in the retention rates of male and female employees. The company can implement focused strategies to improve employee satisfaction, and effective retention for all employees, despite gender, by solving the underlying issues causing the disparity in retention rates.



Active Male Employees: 2475
Inactive Male Employees: 2835

Active Female Employees: 307
Inactive Female Employees: 133

Based on the calculated retention rates:

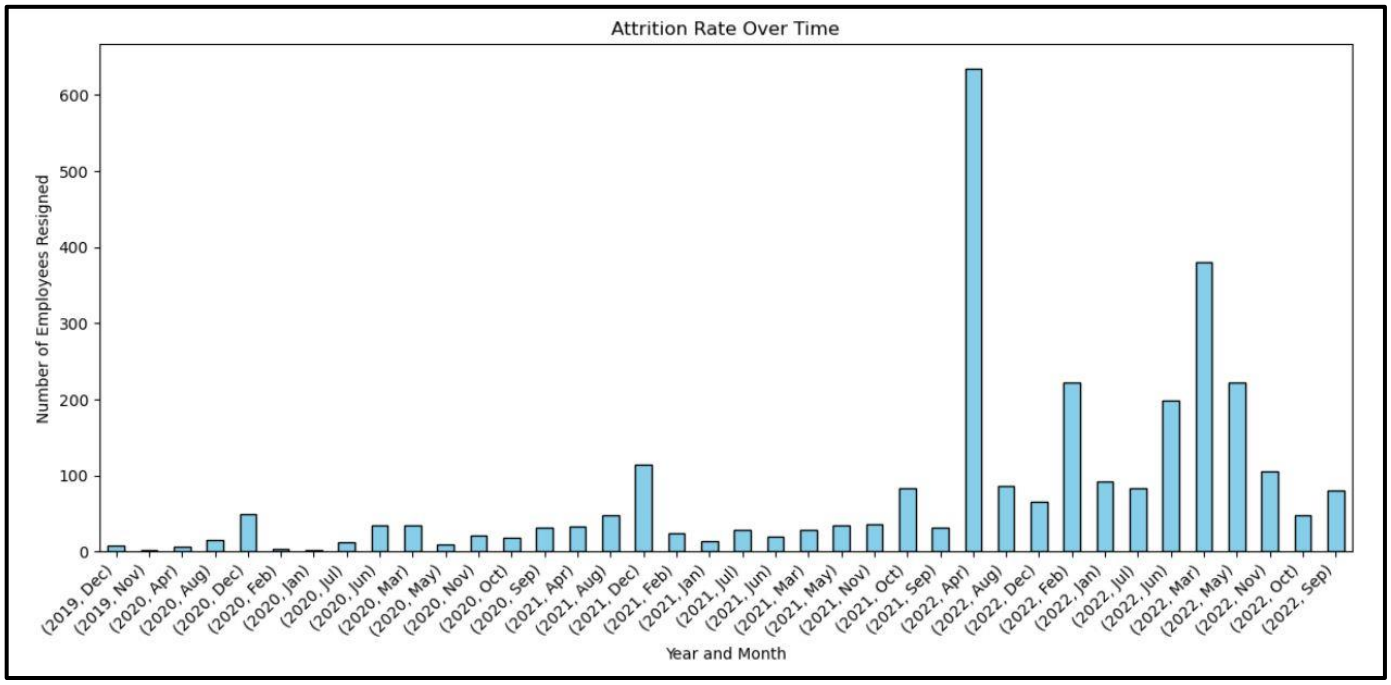
=> The retention rate for male employees is approximately 46.61%.

=> The retention rate for female employees is approximately 69.77%.

The main finding from these retention rates is that there seems to be a big variation in employee retention between male and female employees. The higher retention rate for female employees indicates more female employees than male employees have decided to remain with the organization.

Insight 2: Relationship of time and attrition

This insight aims to understand the rate at which employees are resigning. By analyzing the "Date_Resigned" and "Status" attributes, we can derive the number of employees who resigned from the company in a certain period. This insight can be used to decide the attrition rate. We can get the idea from a bar plot given below.



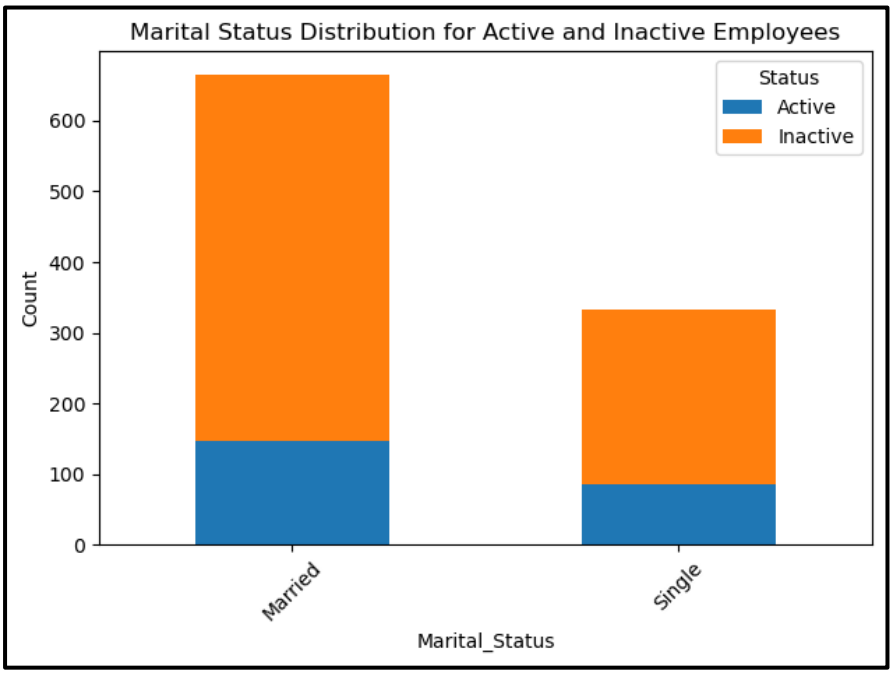
From the plot, we can find,

1. Seasonal pattern: There might be trends influencing attrition rates, as evident from the fluctuating pattern. For example, employee resignations may be higher during certain months of the year and lower during others. For example, **December** month has comparatively higher attritions than other months of the year.

2. Unusual pattern: The presence of outliers, such as **635** in **April** and **381** in **March** resignations in the year **2022**, could be indicative of exceptional circumstances or events that had a substantial impact on employee retention during those periods.

Insight 3: Effect of marital status on retention

To analyze the effect of marital status on employee retention, we can visualize the number of active and inactive employees based on their marital status from a bar plot. From this, we can get the idea of marital status and retention of the employees.



Married Total: 665
Single Total: 332

Married Active: 147
Single Active: 86

The retention rates for each marital status category are as follows:

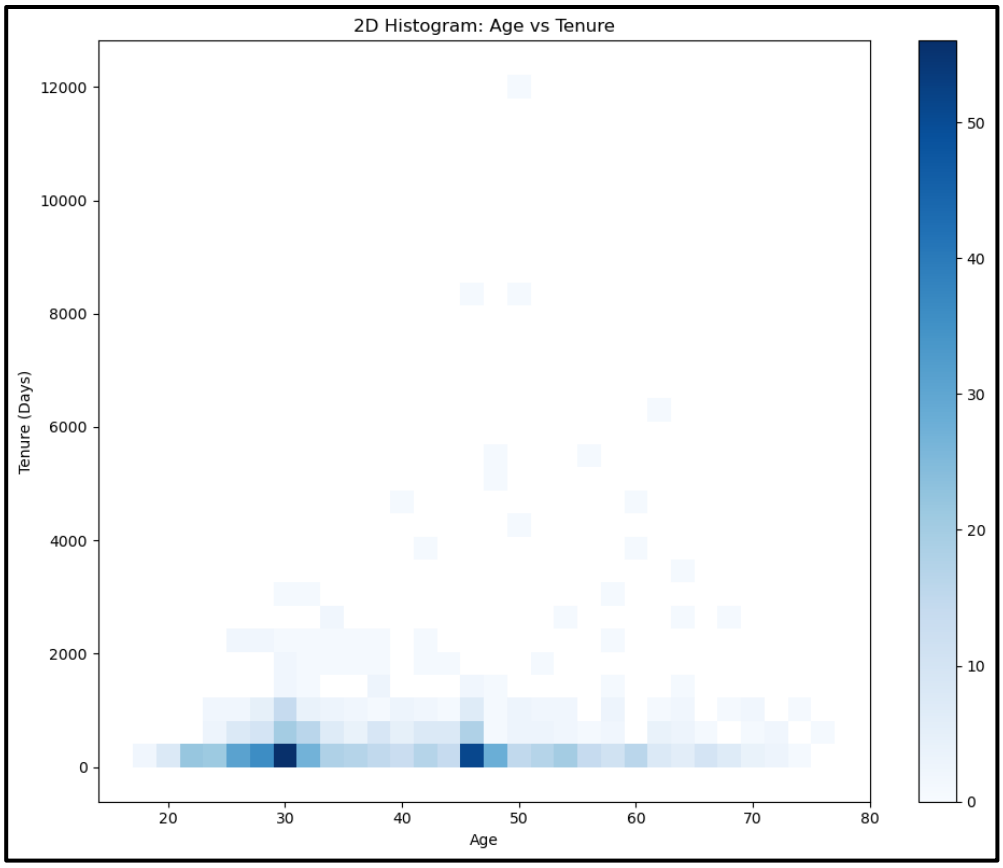
=> Retention Rate for Married Employees: 22.11%

=> Retention Rate for Single Employees: 25.90%

These retention rates indicate the percentage of active employees in each marital status category. From the percentages, single employees' retention rate is a little high than married. So, single employees stay longer than married ones in the organization. But the difference is quite small even though noticeable compared with other factors related to retention.

Insight 4: Relationship between age and tenure

In this analysis exploration of the relationship between employees' Age and their Tenure in the company is possible. The goal of the analysis is to find any correlation between the variables and the likelihood of attrition. From a 2D histogram of age vs tenure, the insight can be derived.



The histogram shows the density of points in various sectors, with darker sectors indicating higher density (more occurrences). From the plot, certain insights were derived.

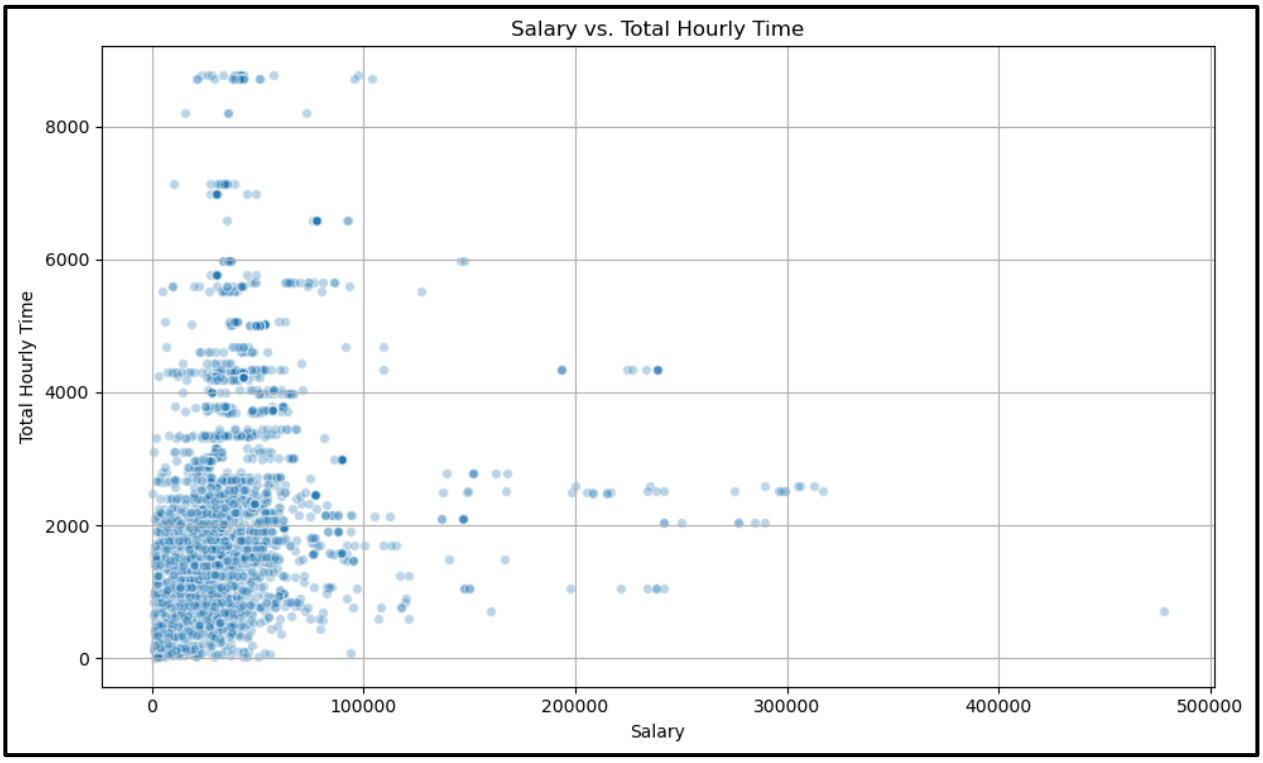
=> Most employees fall in the age range of 25 to 35 and have a tenure of fewer than 1000 days, which means a short retention time compared to other age groups.

=> There is a noticeable number of employees with shorter tenure (around 100 to 300 days) across various age groups.

And finally, the longer-serving employees tend to be older, as shown by the progressive shift towards older age groups as tenure grows.

Insight 5 Relationship between performance and salary

From exploring the relationship between the performance of the employee (measured by factors like attendance) and their salaries. A scatter plot can be used to analyze this, encouraging us to understand the correlation between salary and total working time.



The Total Hourly Time is obtained from the attendance file of each employee and the Salary is derived using the salary file assuming that the Net Salary will represent the salary of an employee.

The correlation coefficient of 0.178 between Salary and Hourly Time is derived.

The scatter plot shows a weak positive correlation between the Hourly Time and their corresponding salaries. While the correlation is positive but not significant.

Therefore, while Hourly Time may be a factor in assessing employee performance, it is essential to consider multiple performance metrics when designing compensation and rewards strategies. This approach ensures that higher-performing employees are adequately rewarded, aligning the organization's compensation structure with overall employee performance and motivation.

Conclusion

In conclusion, the analysis provided important new insight into attrition, performance, and retention of employees. Younger workers with shorter employment histories have been found to be more inclined to depart the company. Additionally, there were gender-based retention disparities that were noticeable, and employee commitment may be somewhat influenced by marital status. Furthermore, a marginally favorable link between salary and job performance was discovered. Gender-related efforts and specialized hiring methods are advised to increase staff retention. continuous analysis of data will contribute to a motivated and engaged workforce.