Project Report

Employee Attrition Analysis and Insights in Marvelous Construction: A Data Science Approach

- **Problem Overview**

Marvelous Construction, a major construction firm operating in Sri Lanka with 35 construction sites, has been experiencing a high rate of employee attrition. The Human Resources department has observed a significant number of employees resigning from the company. This alarming situation has prompted the CEO to seek the expertise of a data scientist to analyze the available data and gain insights into the underlying causes of employee attrition.

Employee attrition can have detrimental effects on an organization, including increased recruitment and training costs, reduced productivity, and a negative impact on employee morale. Understanding the factors contributing to employee attrition is crucial for Marvelous Construction to develop effective strategies and initiatives to improve employee retention and maintain a stable workforce.

- **Dataset Description**

1.employee.csv (631 records):

- Employee_No, Employee_Code, Name, Title, Year_of_Birth, Gender, Religion_ID, Marital_Status, Designation_ID, Date_Joined, Date_Resigned, Reporting_emp_1, Reporting_emp_2, Employment_Category, Employment_Type, Religion, Designation.
- Contains employee details, including join date and resign date for attrition prediction.

2.leaves.csv (237 records):

- Employee_No, leave_date, Type(Half day/Full Day), Applied Date, Remarks, apply_type(Annual/Casual).
- Contains information about employee leaves, with differences observed between new and old employees.

3.salary.csv (2632 records):

- Employee_No, Amount, month, year, <<different factor names>>.
- Includes monthly breakdown of salary, including additions and deductions.

4.attendance.csv (60354 records):

- id, project_code, date, out_date, employee_no, in_time, out_time, Hourly_Time, Shift_Start, Shift_End.
- Provides attendance details, allowing calculation of late minutes by subtracting in time from shift start time.

- **Data pre-processing**

1. ***Data Cleaning***

- ***Missing Values Handling:***
  - Missing values in "Year_of_Birth" are converted to NaN and then imputed using a decision tree-based approach.
  - Missing values in "Marital_Status" are encoded as NaN and imputed using another decision tree-based approach.
    (Note - Handle Complexity in Missing Value Imputation: The decision tree-based approach used for imputing missing values may introduce model complexity and increase running time. If the decision tree model has a large depth or is overly complex, it can become computationally expensive, particularly with a considerable amount of missing data. This might lead to longer processing times and potential performance issues if the model exceeds available resources.
    Solution for this issue: Limit Decision Tree Depth, Parallel Processing, Reduce Dataset Size or KNN efficient approach)

- ***Encoding Categorical Variables:***
  - "Marital_Status," "Gender," and "Employment_Type" are encoded numerically to facilitate analysis.

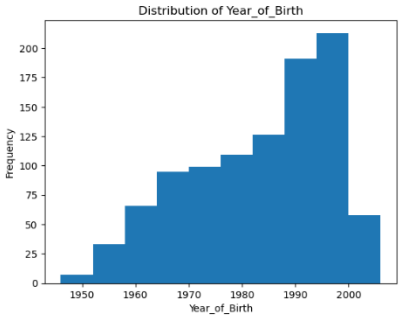
- ***Feature Engineering:***
  - "Date_Joined" is converted to a datetime format and then transformed into the number of days since the minimum date as a new feature "Date_Joined_NumDays."

- ***Handling Missing Values in "Date_Resigned" and "Inactive_Date":***

- To handle missing values in "Date_Resigned" and "Inactive_Date," the code replaces '0000-00-00' with corresponding values from the other column, and '\N' is used to signify missing data more explicitly.

- ***Handling Outliers:***

Outliers are checked and handled using the Z-score method. For each numerical attribute, the upper and lower bounds are calculated based on the mean and standard deviation multiplied by the specified threshold. Any data point outside these bounds is clipped to the nearest bound to mitigate the effect of outliers. It didn't provide any visible insights or conclusions that could be directly used for analysis.

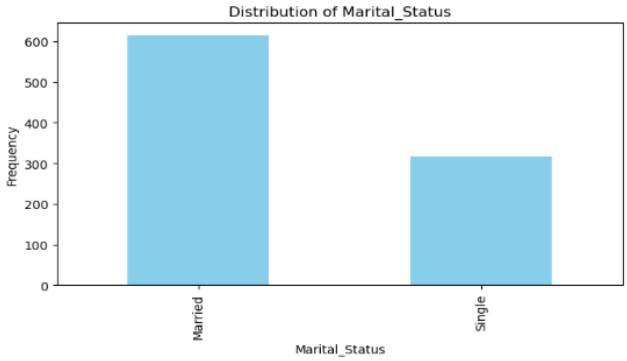


- **Data Analysis**

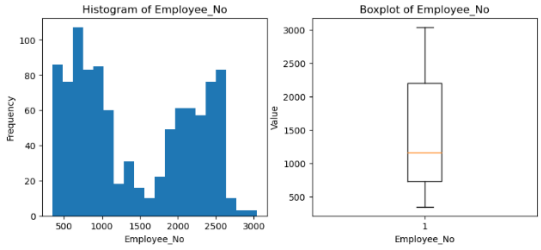
- *Descriptive analysis*

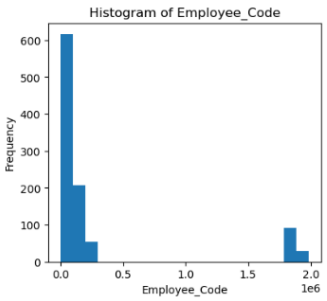
```
Male      932
Female     65
Name: Gender, dtype: int64
Buddhist    842
Hindu       101
Muslim       28
Catholic     26
Name: Religion, dtype: int64
Married    614
Single     316
Name: Marital_Status, dtype: int64
Contarct Basis    977
Permanant          20
Name: Employment_Type, dtype: int64
Inactive    764
Active      233
Name: Status, dtype: int64
```

	Employee_No	Employee_Code	Religion_ID	Designation_ID
count	997.000000	9.970000e+02	997.000000	997.000000
mean	1433.159478	2.794296e+05	1.393180	48.504514
std	770.821006	5.810717e+05	0.967272	43.112201
min	347.000000	6.000000e+00	1.000000	1.000000
25%	728.000000	4.913000e+03	1.000000	27.000000
50%	1164.000000	5.387000e+03	1.000000	31.000000
75%	2197.000000	1.961360e+05	1.000000	55.000000
max	3041.000000	1.985024e+06	5.000000	201.000000



- *Exploratory Analysis*



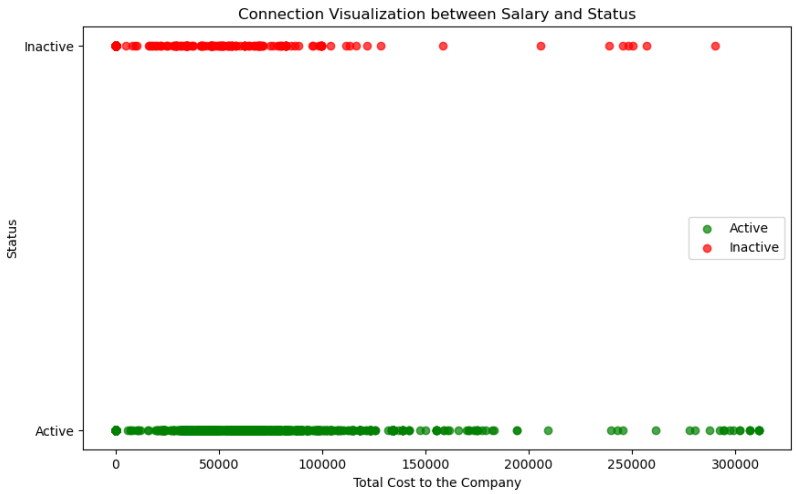


- *Predictive analysis*
  important features for predicting employee status (Active/Inactive) are 'Employee_No', 'Designation_ID', 'gender_Encoded', 'Date_Joined_NumDays', 'Marital_Status_Encoded', 'employment_type_Encoded', and 'Year_of_Birth',
  Accuracy: 0.75
  Precision: 0.77
  Recall: 0.75
  F1-Score: 0.76

- **Insights from data analysis**

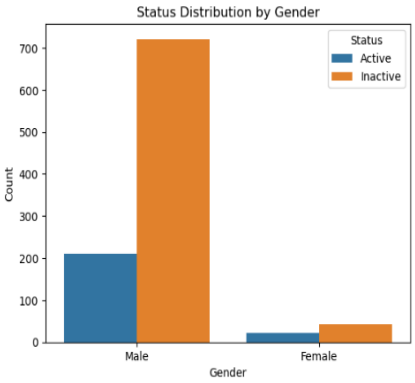
Note: Here active mentions that the employee's status is active at the same time. inactive, meaning the same.

- ***Insight 1****:* Employees who are currently active have higher total cost to the company, which includes their net salary and benefits, compared to inactive employees, suggesting that the overall compensation package may play a role in the reasons for employee resigning.
- Approach: The data of employees' total cost to the company, including net salary and benefits, was collected and categorized based on their status (active or inactive).

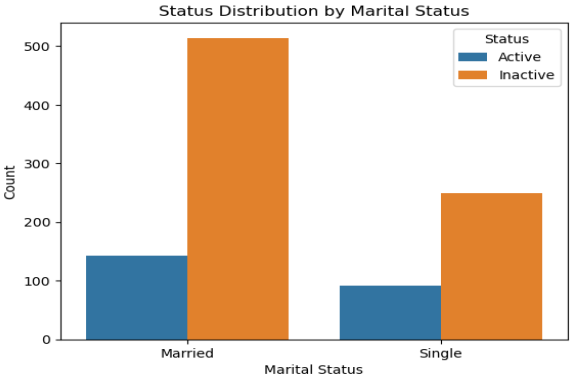


- Analysis: The scatter plot revealed that active employees tend to have higher total cost to the company than inactive employees, indicating that the overall compensation package, which includes net salary and benefits, may influence the reasons for employee inactivity.

- ***Insight 2****:* higher proportion of inactive status among male employees and married employees, suggesting a potential link to resignations, while the count of female employees and single employees is considerably lower.

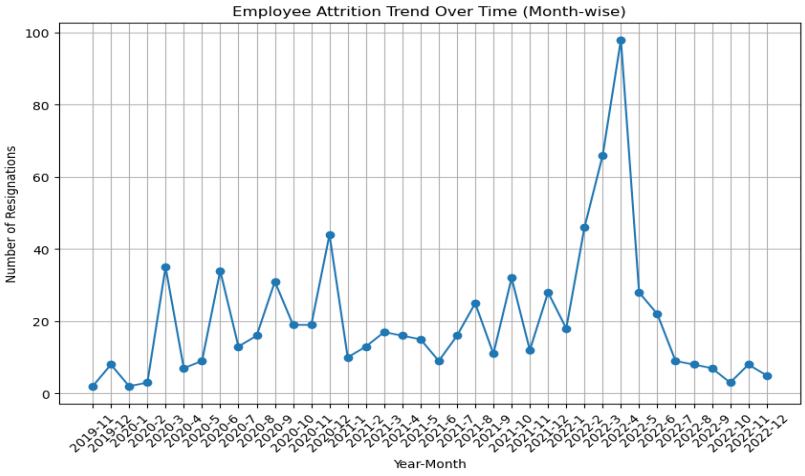


➢ Approach: Utilizing data visualization techniques and exploratory analysis, we investigated the relationship between employee status (active/inactive) and several demographic features, such as gender and marital status.



➢ Analysis: The visualizations revealed a noteworthy pattern: a higher proportion of inactive status was observed among male employees and married employees. This observation suggests a potential link between gender, marital status, and employee resignations. Additionally, the count of female employees and single employees was found to be considerably lower, indicating possible differences in turnover rates based on these demographics.

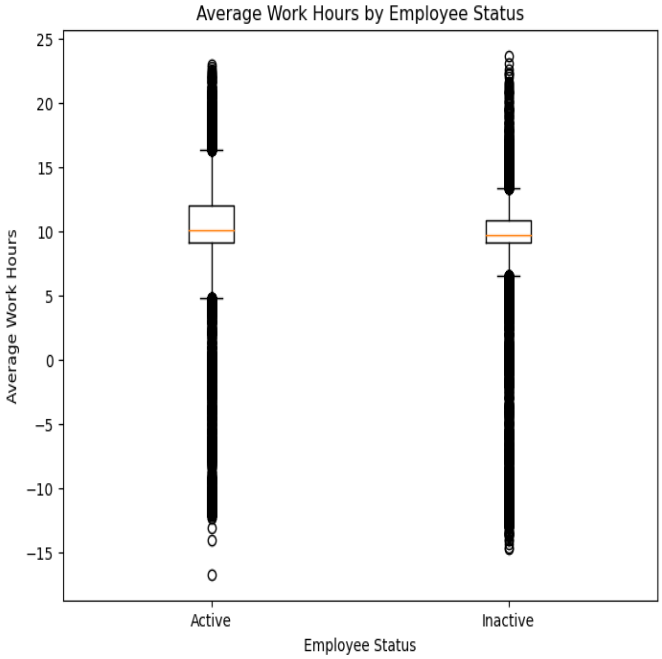
- ***Insight 3***: In 2020 and April 2022, the number of employee resignations was higher compared to other years, while there was a reduction in resignations during 2021.
  - ➢ Approach: Conducting month-wise time-based analysis to understand employee attrition trends over different months of the years.



➢ By tracking attrition rates over each month, the CEO can identify patterns, specific months with higher attrition rates. For instance, certain months might show increased attrition due to factors like holiday seasons, performance review cycles, which can help in implementing targeted retention strategies during those critical periods.

- ***Insight 4:*** Active employees work significantly more hours on average compared to inactive employees.

➢ Approach: We used box plots to visualize and compare the average working hours of two employee groups: Active and Inactive Employees. The box plots were created side by side, with the x-axis representing the employee groups and the y-axis representing the working hours.



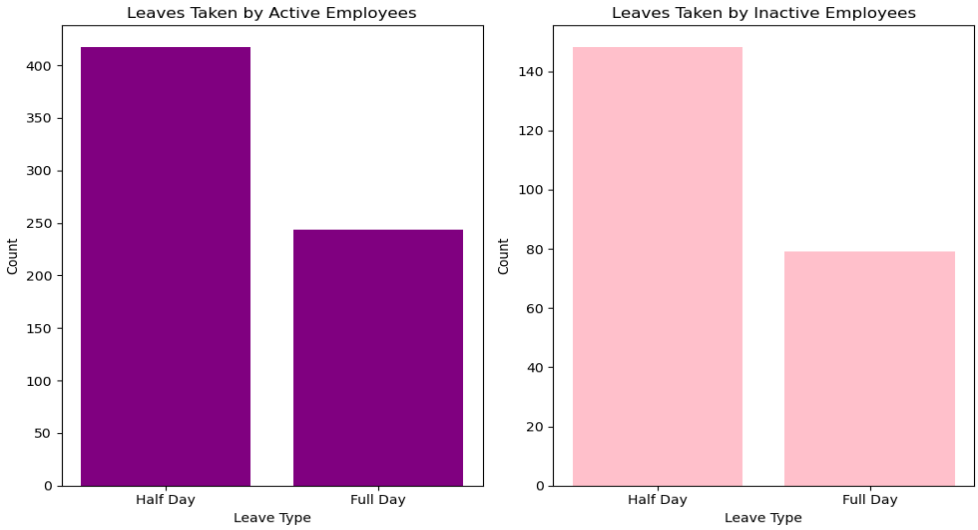
➢ Analysis:

The box plot for Active Employees showed a higher median working hour value, indicating that the typical working hours for Active Employees are longer than for Inactive Employees. Inactive Employees generally work fewer hours on average compared to Active Employees.

Spread of Data: The box plot for Active Employees had a wider box, indicating more variability in their working hours. On the other hand, the box plot for inactive employees had a narrower box, implying less variation in their working hours. This suggests that Inactive Employees tend to have more consistent working schedules.

- ***Insight 5:*** Active employees take more leaves overall than inactive employees, while both groups primarily prefer taking half-day leaves.

➢ Approach: The approach taken for this analysis involves examining the leave patterns of employees and their correlation with employee status and other attributes. The first step is to merge the data from the "leaves.csv" and "employees.csv" files using the common 'Employee_No' column. Then, we segment the data into two groups based on employee status: active and inactive employees.



- Analysis: After segmenting the data, we visualize the distribution of leaves taken by active and inactive employees using bar plots. The analysis reveals that active employees tend to take more leaves than inactive employees. Surprisingly, both groups primarily prefer taking half-day leaves over full-day leaves. This insight suggests that employees, regardless of their status, value work-life balance and often choose to take shorter leaves to maintain productivity while fulfilling personal needs. These findings could be valuable for designing leave policies that prioritize employee well-being, leading to improved job satisfaction and potentially lower attrition rates.