CS3121 Introduction to Data Science

Final Project Report

• Problem overview

Marvelous Construction is a major construction firm with 35 construction sites in different areas in Sri Lanka. There seems to be a higher attrition rate in employees in the company which was noted by the Human Resource Department of the organization.
It is needed to analyze the data within the company to analyze the situation and find out reason for employees leaving in a large count.

• Dataset description

The dataset contains employee details, attendance, leaves, and salary extracted from the ERP of Marvelous Construction.
More details about each data file are given below.

File Name	# Records	Dimension	Remarks
employee	997	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	Join date, Resigned date are special fields for training and testing data set for attrition prediction
leaves	1018	Employee_No, leave_date, Type (Half day/Full Day), Applied Date, Remarks, apply_type(Annual/Casual)	New employees have fewer leaves, while old employees have a higher no of leaves.
salary	9035	Employee_No, Amount, month, year, <<different factor names>>	Monthly addition/deduction breakdown is included
attendance	224057	id, project_code, date, out_date, employee_no, in_time, out_time, Hourly_Time, Shift_Start, Shift_End	Late minutes = in time - shift start time

❖ Data pre-processing on employees.csv

➢ Handling missing values in Year_of_Birth and Marital_Status simultaneously.

- The Year_of_Birth column is converted to numeric data type. Any values equal to ‘0000‘ are replaced with NaN, which will be considered as missing values.

1. For rows with only Year_of_Birth missing.

    I. The mode of Year_of_Birth is calculated separately for rows where the Marital_Status is "Single" and "Married".

    II. Missing values in the Year_of_Birth column are filled based on the Marital_Status. For rows where Marital_Status is "Single" and Year_of_Birth is null, it is filled with the mode value calculated for singles. Similarly, for rows where Marital_Status is "Married" and Year_of_Birth is null, it is filled with the mode value calculated for married individuals.

2. For rows with only Marital_Status missing

    I. The data is grouped by Year_of_Birth and the mode of Marital_Status is selected within each group.

    II. Missing values in the Marital_Status column are filled using the mode values calculated in the previous step. The Year_of_Birth values are mapped to the corresponding mode values.

3. Filling Missing Values in Marital_Status and Year_of_Birth When Both Are Missing

    - For rows where both Marital_Status and Year_of_Birth are missing, missing values in Marital_Status are filled with the mode of the entire column. Missing values in Year_of_Birth are filled with the integer value of the median of the column.

- **Updating invalid and missing Inactive_Date and Date_Resigned Values**

1. If the status is Active the employee can’t have a valid Inactive_Date or valid Date_resigned.
   So, when Status is ‘Active’, if Inactive_Date is "0000-00-00”, changed it to '\N'.
   Similarly, when Status is ‘Active’, if 'Date_Resigned is "0000-00-00", change it to '\N'.

2. If Inactive_Date is not "0000-00-00" and Date_Resigned is "0000-00-00", replace the "0000-00-00" in Date_Resigned with the corresponding Inactive_Date .

3. If Date_Resigned is not "0000-00-00" and Inactive_Date is "0000-00-00", replace the "0000-00-00" in Inactive_Date with the corresponding Date_Resigned .

- **Updating invalid titles of Employees**

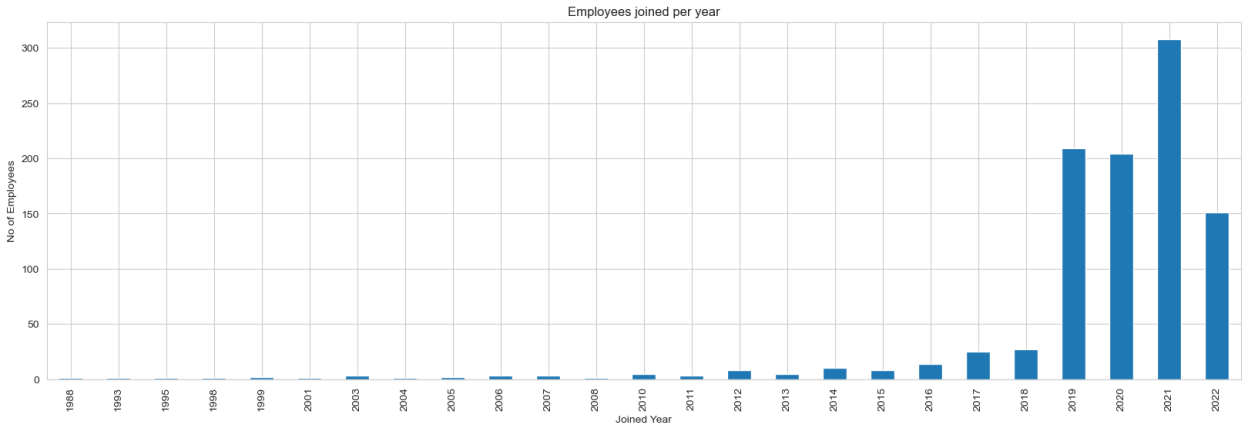
  - Titles are updated for employees. For male employees with titles 'Ms' or 'Miss', the title is changed to 'Mr'. For female employees with the title 'Mr', a random choice between 'Ms' or 'Miss' is assigned.

- **Updating invalid 'Reporting_emp_1' Values**

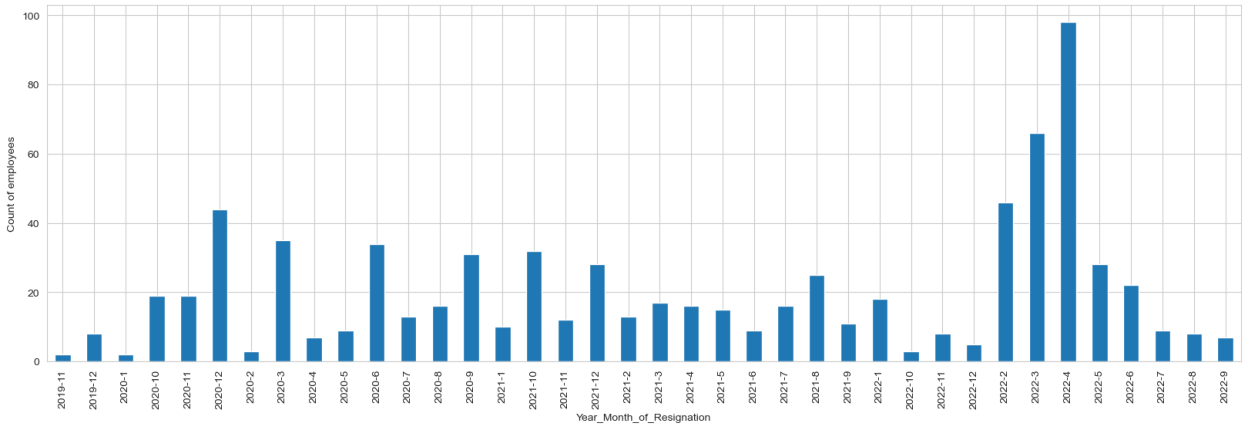
  - If 'Reporting_emp_1' is the same as 'Employee_No', it is changed to '\N'.
    Because an employee won’t be reporting to their self.

• Insights from data analysis

1. This is a bar plot showing the Employees joined per year.
   And most of the employees (more than 800) seems to have been joined in years 2019, 2020, 2021, 2022.



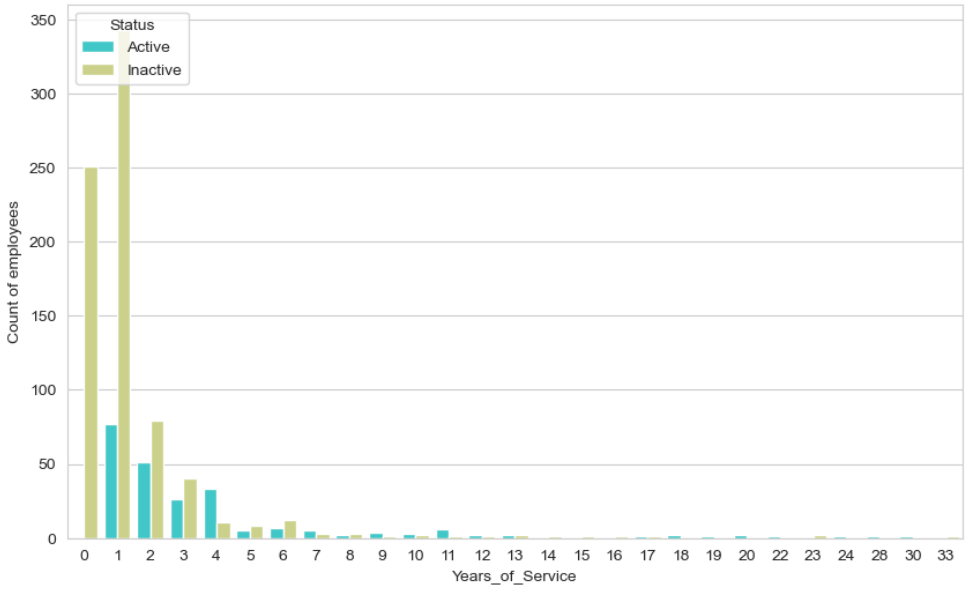
2. The below bar plot shows the count of employees resigned for each month of years from 2019 to 2022. Only the months with employees resigned are shown here.



From October 2020, mostly every month, there seems to be employees leaving the organization.

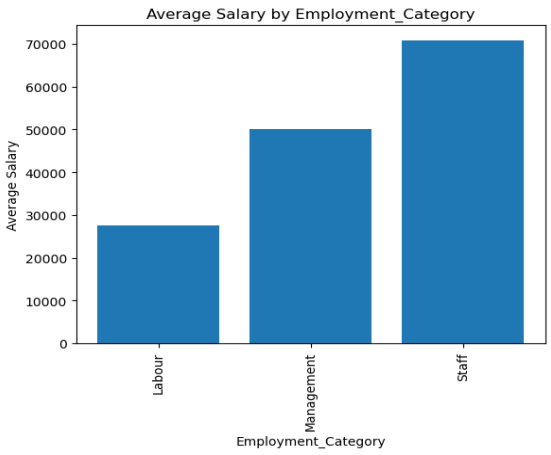
In 2022, from February to April, for consecutive of three months, the attrition rate is higher and a total of over 200 employees resigned.

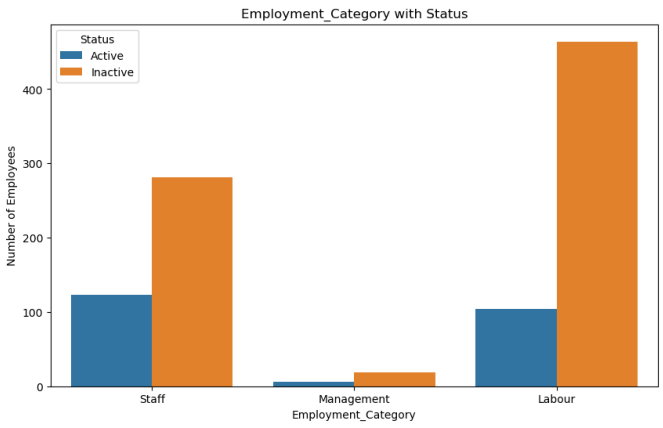
3. Most of the employees seems to resign before completing atleast 2 years of service.



And the Employees joined per year graph (from insight 1) shows most employees joind in the last 4 years. It can be said employees joined and left in the last 3-4 years in a high number.

4. The Average Salary by Employment_Category shows average salary of a labor is very less compared to Staff or the management employees.

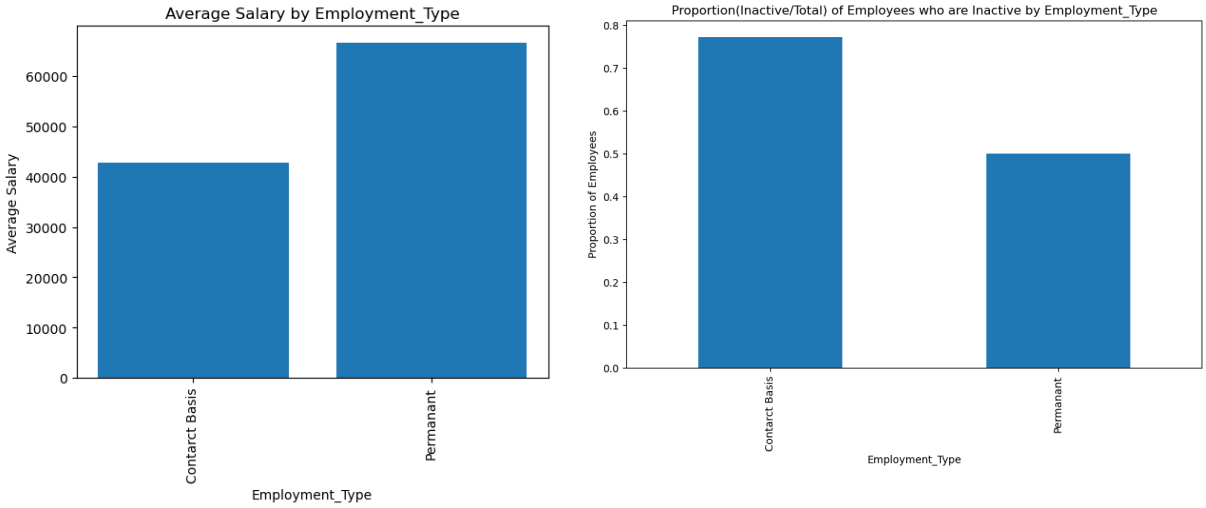




And the Employement_Category with Status graph shows that with in an

employment_Category labors are the one who seems to be at higher proportion of resigning their job.

5. Contract Base employees have lower salary compared to Permanent employees. And They have the higher proportion of resigning the the job with in their Employment_Type as the 'Proportion(Inactive/Total) of Employees who are Inactive by Employment_Type' graph shows.



Moreover Contract Basis empoyees resigning have more impact on attrition rate as because they are 98% of total employees.