CS3121 - Introduction to Data Science

Employee Attrition in Marvelous Construction

Project Report - Group 08

210049U | 210086E | 210173T | 210572P | 210574A

1. Problem Overview

Marvelous Construction, a major construction firm in Sri Lanka with 35 construction sites, is experiencing a high rate of employee resignations. This trend has raised concerns within the Human Resources department and prompted the CEO to hire a data scientist to analyze internal data and understand the situation better. The goal of this project is to analyze the provided dataset, which includes employee details, attendance, leaves, and salary information extracted from the company's ERP system. The objective is to derive actionable insights that will help the CEO make strategic decisions to improve employee retention.

2. Dataset Description

This dataset includes:

- employee.csv - All employee details
- leaves.csv - Details about the leaves taken by each employee
- salary.csv - Details about the salary of each employee
- attendance.csv - Details about employee attendance
- salary_dictionary.csv - Different types of salaries and cuts

The main focus is on the employee file, which contains employee names, IDs, religions, genders, dates of birth, employment types, employment categories, designations, and other relevant employment details. This file is crucial for preprocessing and analysis.

3. Data pre-processing

Data preprocessing is a crucial step in machine learning where raw data is transformed into a clean, structured format suitable for analysis and modeling. This process involves various techniques such as handling missing data, encoding categorical variables, and scaling features. The goal is to prepare the data in a way that allows machine learning algorithms to learn patterns effectively and make accurate predictions.

- **Unique Value Identification:**

  Using the unique() function simplifies this process by extracting all unique values from a categorical column. These unique values can then be used for filtering, grouping, or further analysis to gain valuable insights from the data

- **Resignation Status Encoding:**

  The approach you've taken to create a new column, 'Is_resigned', based on the presence of values in the 'Date_Resigned' column is a common and practical way to handle such situations. By assigning 0

to rows where 'Date_Resigned' is null and 1 to rows where it is not null, you effectively create a binary indicator variable that represents whether an employee has resigned or not.
Directly mapping 1 and 0 values to the 'Date_Resigned' column could lead to misinterpretation or errors in analysis, especially if there are missing values or if the dates are not consistently formatted. By creating a separate column, you maintain the integrity of the original 'Date_Resigned' column and provide a clear and easily interpretable indicator of resignation status.

- **String to Datetime Conversion:**

The technique used to convert date strings in the format '02/05/2023' to datetime objects. This process involves using the pd.to_datetime() function provided by the pandas library to convert date strings into a standardized datetime format.

- **Dates Features Engineering:**

various date-related features were extracted from the 'Date_Joined', 'Date_Resigned', and 'Inactive_Date' columns in the employee dataset. This included creating new columns for the year, month, and day of joining, resignation, and inactivity. Missing dates were filled with a placeholder date to ensure consistency. The new columns were converted to integer data type for easier analysis. This feature engineering step enhances the dataset, providing additional insights for analyzing hiring patterns, attrition, and inactivity trends.

- **Missing Value Analysis and Column Removal:**

The missing value analysis was conducted to identify columns with missing values. The number of missing values for each column was calculated and displayed. Columns with a high number of missing values were identified and deemed unnecessary for analysis. As a result, the 'Reporting_emp_1' and 'Reporting_emp_2' columns were removed from the dataset to streamline further analysis. This process ensures that the dataset is clean and contains only relevant columns for the analysis, reducing noise and improving the quality of the dataset.

- **Irrelevant Column Removal:**

In addition, further irrelevant columns were identified and dropped from the dataset. Columns such as 'Employee_Code', 'Name', 'Religion_ID', 'Designation_ID', 'Date_Joined', 'Date_Resigned', and 'Inactive_Date' were deemed to have no direct impact on the final output or analysis. These columns were therefore removed to simplify the dataset and focus on relevant features for the analysis. This step ensures that only meaningful and impactful features are retained, reducing complexity and improving the efficiency of the analysis process.

- **One-Hot Encoding:**

To handle categorical variables like 'Title' and 'Religion', which are not ordinal, one-hot encoding was applied. This technique converts categorical variables into binary vectors, creating new columns for each category and assigning a 1 or 0 to indicate the presence or absence of each category in the

original column. This process ensures that the model does not interpret categorical variables as ordinal, preserving the integrity of the data and improving the accuracy of the analysis.

- **Binary Categorical Variable Conversion:**

Binary categorical variables ('Gender', 'Marital_Status', 'Status', 'Employment_Type') were converted to numeric values (1s and 0s) using a predefined mapping. This mapping assigned 1 to categories such as 'Male', 'Married', 'Active', and 'Permanent', and 0 to their respective counterparts ('Female', 'Single', 'Inactive', 'Contract Basis'). This conversion simplifies the representation of binary categories, making them more suitable for analysis and modeling. Additionally, the columns were renamed to reflect their converted nature ('Is_Male', 'Is_Married', 'Is_Active', 'Is_Permanent'), providing clearer column names for future reference.

- **Ordinal Categorical Variable Mapping:**

The 'Employment_Category' column, representing ordinal categorical data with categories like 'Management', 'Staff', and 'Labour', was mapped to integer values using a predefined mapping. This mapping assigns a numerical value to each category based on its ordinal position, with 'Management' being the highest category (3), 'Staff' the middle category (2), and 'Labour' the lowest category (1). This mapping allows for the representation of ordinal data in a way that preserves the inherent order of the categories, making it suitable for analysis and modeling.

- **Adding Salary and Attendance Information to Employee Dataset:**

Salary information was added to the employee dataset ('employees') by calculating the mean basic and net salaries for each employee based on the 'Employee_No' column in the 'salary_df' dataset. Rows with basic salaries of 0 were filtered out to ensure accuracy. The mean salaries were rounded to two decimal places and merged into the 'employees' dataset. The column name 'Basic Salary_0' was renamed to 'Basic Salary' for clarity. This process enriches the dataset with salary information, which can be valuable for further analysis and decision-making.

Attendance information was added to the employee dataset ('employees_with_salary') by first removing outlier values from the 'attendance_df' dataset. The 'date' column was converted to datetime format, and only records up to '2023-12-31' were retained. Next, the number of attendance records for each employee was calculated and stored in 'attendance_unique_values'. This count represents the number of times each employee appeared in the attendance records. Finally, the attendance information was merged into the employee dataset based on the 'Employee_No' column. This process enhances the dataset with attendance data, providing insights into employee attendance patterns.

- **Visualization of Null Values Using Heatmap:**

A heatmap was created to visualize the presence of null values in the 'employees_with_attendance' dataset. The heatmap allows for easy identification of columns with missing values, represented by

the absence of color (e.g., green) for non-null values and the presence of color for null values. This visualization aids in understanding the completeness of the dataset and identifying areas that may require further data cleaning or imputation.

- **Salary Analysis by Designation Category:**

This part analyzes salaries based on designation categories in the 'employees_with_attendance_copy' dataset. It iterates through each unique designation category, filters the dataset for that category, and plots the salaries of employees in that category. Categories with more than 10 non-null values in the 'Basic Salary' column are considered for analysis. For each category, the plot includes a scatter plot of employee salaries, the mean salary (red dashed line), and the median salary (green dashed line). This analysis provides insights into salary distributions within different designation categories, helping to identify trends and anomalies in salary data.

- **Category-Wise Median Imputation:**

The function `categories_median_imputation` is designed to impute missing values in the 'Basic Salary' and 'Net Salary' columns based on category-wise medians. It calculates the global median salary for the entire dataset and then computes category-wise medians for each designation category. If a designation category has more than the specified threshold number of non-null values (e.g., 10), the category-wise median is used for imputation; otherwise, the global median is used. The function iterates through the dataset and updates the missing values in the specified columns with the imputed values. This imputation strategy ensures that missing salary values are filled with values that are representative of each designation category's salary distribution.

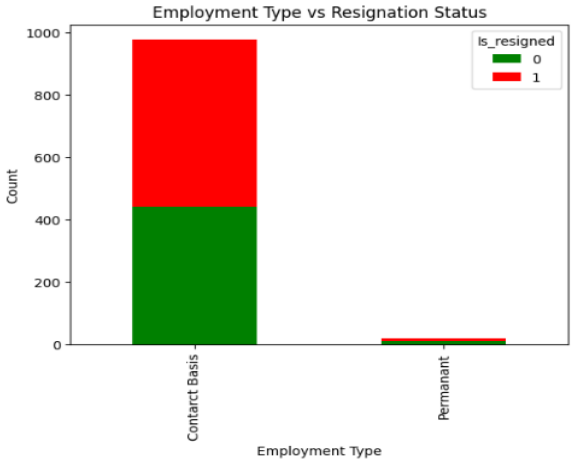
- **KNN Imputation for Missing Values in Year of Birth and Marital Status Dataset:**

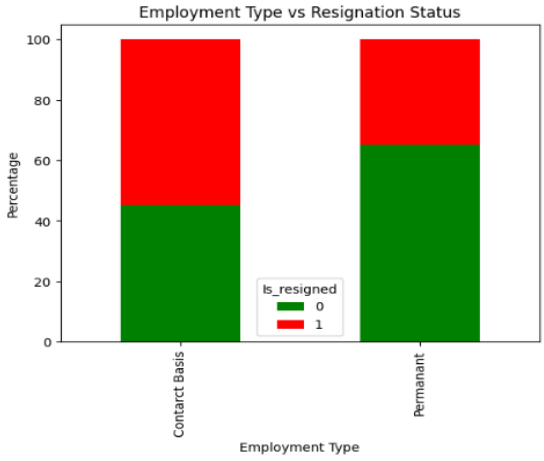
This part demonstrates the implementation of K-Nearest Neighbors (KNN) imputation to fill missing values in the 'employee_yob' dataset, excluding the 'Employee_No' column. The KNNImputer from the sklearn.impute module is used with n_neighbors=5 to impute missing values based on the mean value of the five nearest neighbors. The imputed data is then converted back to a DataFrame, and the 'Employee_No' column is added back to the dataset. Finally, the code prints the number of missing values in each column to confirm that all missing values have been successfully imputed.

We use K-Nearest Neighbors (KNN) imputation to fill missing values in the 'employee_married' dataset, excluding the 'Employee_No' column. The KNNImputer from the sklearn.impute module is employed with n_neighbors=33 to impute missing values based on the mean value of the 33 nearest neighbors. The imputed data is then converted back to a DataFrame, and the 'Employee_No' column is reintegrated into the dataset. Finally, the code prints the number of missing values in each column to verify that all missing values have been successfully imputed

4. Insights from data analysis

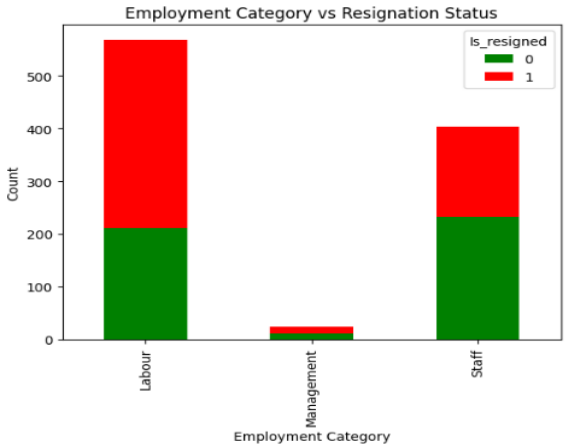
1) Contract basis employees are more likely to resign compared to permanent employees.

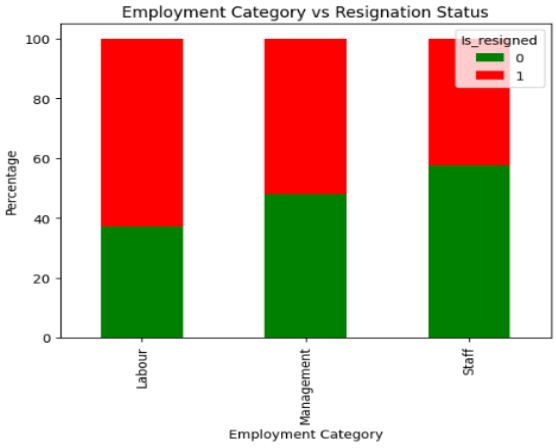




It seems that a considerable number of employees on a contract basis have quit their jobs. This can be seen from both the count and percentage graphs. On the other hand, even though the total number of permanent employees is lesser, their resignation rate and number of resignations are lower than that of employees on a contractual basis.

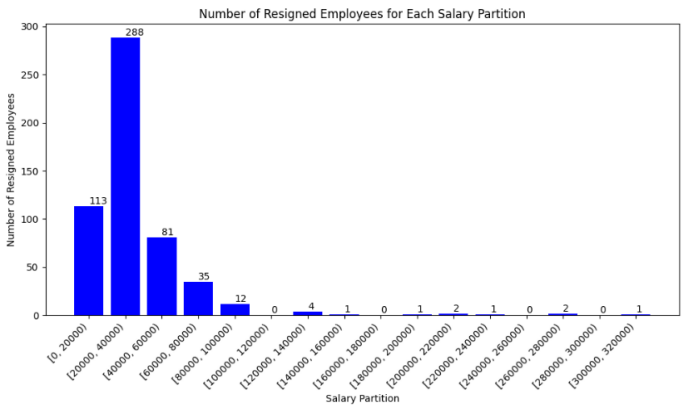
2) The Labour category employee has a higher probability of resigning and accounting for the highest percentage of the overall population

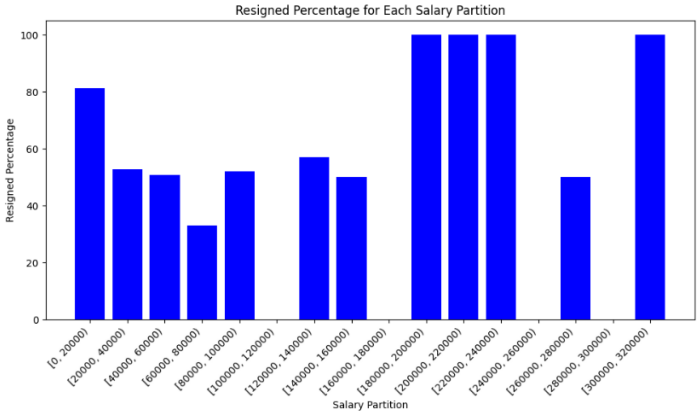




The number of individuals in the management category is very low, accounting for only a small percentage of the overall population. However, a significantly higher percentage of individuals in this group have resigned. This may indicate that there are fewer employees in management roles and that turnover in these positions is relatively high. On the other hand, the staff category has a moderate count and percentage of individuals, with the lowest portion of individuals having resigned. This suggests that staff roles may have a lower turnover rate compared to labor and management roles.

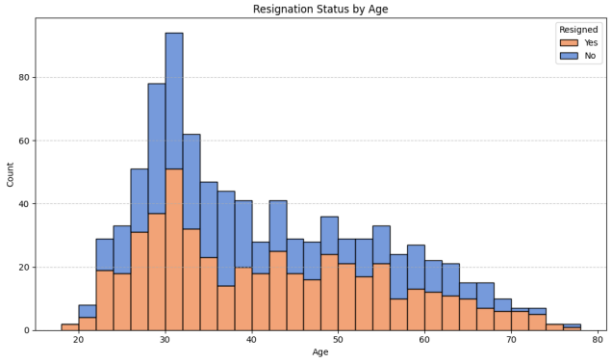
3) The resignation rates vary significantly among different salary partitions. The highest number of employees well as the highest number of resignations are in the '20000-39999' salary bracket, and there are noticeable peaks in resignation percentages at certain salary partitions

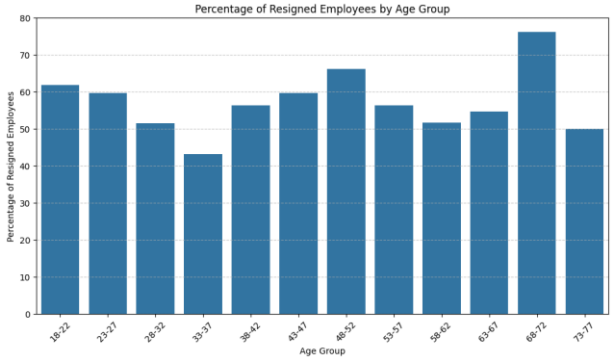




It has been observed that employees in higher salary brackets are more likely to resign than those in lower salary brackets. It is worth noting that even if there are only one or two employees in the higher salary bracket, it can significantly affect the overall resignation rate. On the other hand, although there are more employees in lower salary brackets, even a low percentage of resignations can result in a significant number of people leaving the company.

4) Age of Employees is not a considerable factor to the resignation rate. But a higher number of resignations reported from age 25-35 employees. But we can not purely decide on this number because the amount of employees currently active is also high.





When we look into the percentage of resigned employees by Age, age between 28-37 employees has the lowest percentage of resignation. Also, at the age above 63 we can have a high variation. This is because there is less data in these ages. So we can only get a rough estimate that employees above 60 are more likely to retire.

5) Employees often leave their jobs within the first few years, especially in the beginning. It's really important to prioritize the satisfaction and fulfillment of new employees.. If they're satisfied with their job early on, they're more likely to stay longer. So, it's crucial to focus on making new employees feel welcome and supported right from the start to keep them around for a while.

5. Results of Hypothesis Testing

Hypothesis testing is a critical statistical method used to assess the validity of claims or assumptions about a population parameter based on sample data. It allows researchers to make informed decisions about whether observed differences or relationships are statistically significant or simply due to chance. In hypothesis testing, two mutually exclusive statements, the null hypothesis (H0) and the alternative hypothesis (Ha), are formulated and tested against each other using appropriate statistical techniques.

- Ho: There is no significant impact of each factor on the resignation of the employees.
- Ha: Each factor significantly impacts the resignation of the employees.

Here are the hypotheses that we choose to test in this analysis:

- H1: Marital status significantly impacts the resignation of employees.
- H2: Employment category significantly impacts the resignation of employees.
- H3: Religion significantly impacts the resignation of employees.
- H4: Net Salary significantly impacts the resignation of employees.
- H5: Attendance significantly impacts the resignation of employees.

To test the hypotheses, we employ the following statistical tests:

- **Chi-Square Test: (H1, H2, H3)**

We employ the Chi-square test to evaluate the significance of associations between various categorical variables and employee resignation. This method is apt for investigating whether factors such as marital status, employment category, and religious affiliation have a discernible impact on the likelihood of employees resigning. By utilizing the Chi-square test, we can ascertain whether there exists a meaningful relationship between these categorical variables and the occurrence of employee resignation, providing valuable insights into the dynamics influencing workforce retention within the organization.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Here are the calculated p-values for each test:

H1: 0.222852563405215
H2: 1.933418669876751 e-09
H3: 0.019948406511455

We choose alpha (significance level) as 0.05. The p-values of *H2 and H3* are smaller than alpha. Therefore we have enough evidence to reject the null hypothesis. So, we can say that ***Employment Category and Religion significantly impact the resignation of the employees***. The p-value of *H1* is greater than alpha. Then there is no evidence to reject the null hypothesis. Therefore, ***there is no significant impact of Marital Status on the resignation of the employees***.

- **One-way ANOVA Test: (H4, H5)**

We utilize the one-way ANOVA test to assess the presence of statistically significant differences in net salary and attendance across diverse groups. This method is particularly appropriate for evaluating the impact of a continuous independent variable on a categorical dependent variable. By employing the one-way ANOVA test, we aim to determine whether variations in net salary and attendance are attributable to genuine differences between groups or are simply the result of random variability.

Here are the calculated p-values for each test:

H4: 0.0004606332619948
H5: 1.5824113770828567 e-27

When the alpha equals 0.05, both p-values are smaller than alpha. So, we have enough evidence to reject the null hypothesis and accept the alternative hypothesis. Therefore, we can say that ***Net Salary and Attendance significantly impact the resignation of the employees***.