Employee Attrition in Marvelous Construction

Final Project Report

1. Problem overview

Marvelous Construction, a major construction firm in Sri Lanka, is facing a concerning issue of high employee attrition across its 35 construction sites. In response, the company has hired a data scientist to analyze the data and provide valuable insights for improving employee retention. The objective of this project is to preprocess the dataset, conduct a thorough analysis, and derive key insights that will empower the CEO to make strategic decisions in addressing the problem effectively.

To ensure the reliability and quality of the dataset, preprocessing steps will be implemented. This involves handling missing values, identifying and rectifying inconsistencies or errors, and transforming variables if necessary. By cleaning the data, we can ensure that the subsequent analysis is based on accurate and dependable information.

Following the preprocessing stage, a comprehensive analysis will be conducted to identify the underlying factors contributing to employee resignations. Key insights will be derived by applying various data analysis techniques, supported by relevant visualizations and in-depth analysis. These insights will shed light on crucial aspects such as job satisfaction, work environment, compensation, career development opportunities, and employee engagement. By understanding these factors, the CEO can make informed decisions to implement targeted strategies and policies aimed at enhancing employee retention.

In summary, this project aims to address the pressing issue of high employee attrition at Marvelous Construction through data analysis. By leveraging the power of data, the CEO will gain valuable insights into the factors influencing attrition and be equipped to make informed decisions. The ultimate goal is to develop effective measures that will mitigate employee turnover, improve overall performance, and drive the success of the company.

2. Dataset Description

The dataset provided for analysis in this project consists of four files:

1. **"Employee,"**
2. **"leaves,"**
3. **"salary"**
4. **"attendance."**

The "employee" file contains 631 records and includes information such as 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, and Designation. Notably, the Date_Joined and Date_Resigned fields will be used for training and testing data sets for attrition prediction.

The "leaves" file consists of 237 records and provides details about employee leaves, including Employee_No, leave_date, Type (Half day/Full Day), Applied Date, Remarks, and apply_type (Annual/Casual).

The "salary" file contains 2632 records and includes information about Employee_No, Amount, month, year, and various factors related to monthly addition and deduction breakdown.

Lastly, the "attendance" file consists of 60,354 records and provides data on employee attendance. It includes fields such as id, project_code, date, out_date, employee_no, in_time, out_time, Hourly_Time, Shift_Start, and Shift_End. Additionally, the "attendance" file also calculates the late minutes by subtracting the in-time from the shift start time.

Overall, this dataset offers a comprehensive view of employee-related information, including personal details, leaves, salary breakdown, and attendance records. By analyzing this dataset, we aim to gain valuable insights into the factors contributing to employee attrition and derive actionable recommendations to improve employee retention at Marvelous Construction.

3. Data pre-processing

A series of preprocessing steps were conducted on the **"employee"** dataset to ensure data quality and prepare it for further analysis. The steps involved handling missing values, transforming categorical variables, calculating additional features, and imputing missing values.

1. **Changing Genders and Marital Status:**

The code updates the "Title" column based on gender and marital status. If the gender is male, the title is changed to "Mr." If the gender is female and the marital status is single, the title is changed to "Miss." If the gender is female and the marital status is married, the title is changed to "Ms."

2. **Handling Missing Values:**

The code identifies and handles missing values in the dataset. It first identifies the rows where the year of birth is "'0000'" and replaces it with "0000" to represent a missing value. The "Date_Joined" column is then converted to datetime format.

3. **Label Encoding**:

Label encoding is applied to categorical columns using the LabelEncoder from scikit-learn. The columns encoded include "Gender," "Status," "Title," "Employment_Category," and "Employment_Type." The encoded values are stored in temporary columns with suffix "_temp."

4. **Age-related Calculations**:

The code calculates the "Age_at_Joining" column by subtracting the year of birth from the year of joining. Additionally, the "Years_of_Service" column is calculated by subtracting the joining year from 2023.

5. **Imputation of Missing Values:**

To impute missing values in the "Year_of_Birth" column, the dataset is split into a train and test dataset based on the availability of the target variable. Random Forest regression is trained on the train dataset using the available features to predict the missing values in the test dataset.

6. **Imputation of Marital Status:**

For the missing values in the "Marital_Status" column, a Random Forest classifier is trained on the available data. The trained model is then used to predict the missing values in the test dataset.

7. **Data Integration:**

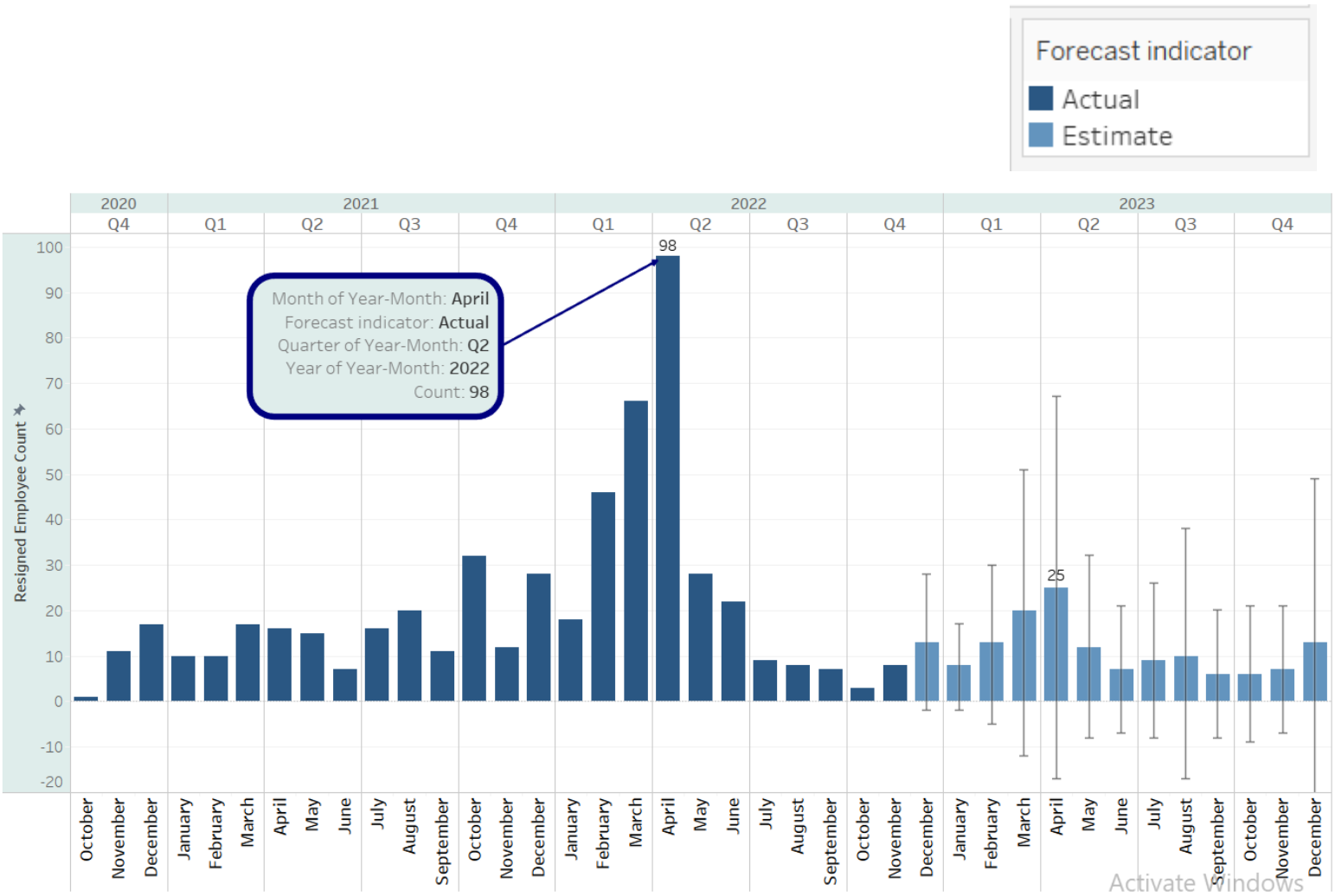
The imputed values for "Year_of_Birth" and "Marital_Status" are integrated back into the original dataset using the Employee_No as a key.

8. **Data Cleaning**:

Finally, the columns in the preprocessed dataset are rearranged to match the original dataset's column order, and the resulting preprocessed dataset is saved as a CSV file.

4. Insights from Data Analysis

Insight 1 - Time Series Analysis of Attrition and Forecast



- According to the above time series analysis we can see that there was a surge in attrition of employees during 1st and 2nd quarters of the fiscal year of 2022.
- During the next two quarters, resignation rate went down
- Given above data we can estimate what would happen in 2023, we can expect another wave of attrition during the 2nd quarter which might not be as extreme as the previous year.
- Hence investigating the decisions made from the administration of the company during the 1st and 2nd quarters would be important to figure out what caused the sudden rise of attrition.

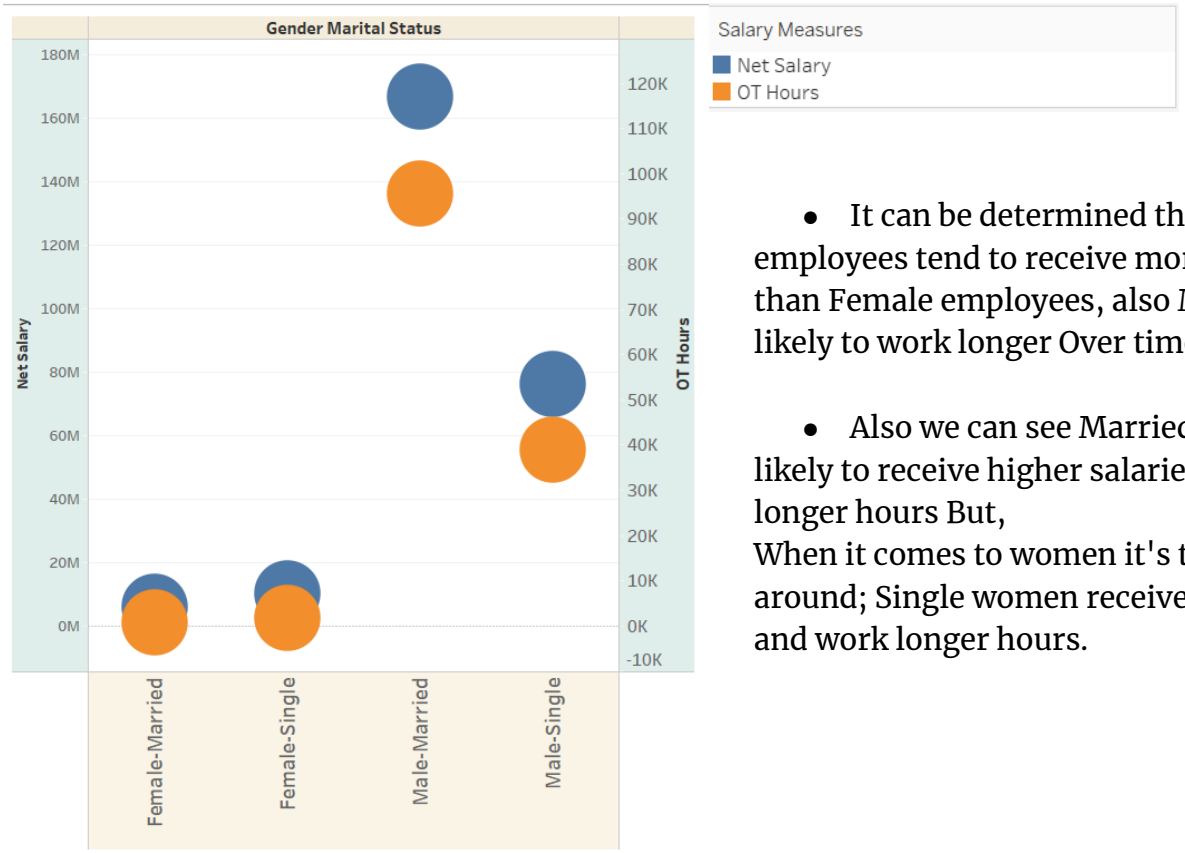
Here are the details about the model used for forecasting.

All forecasts were computed using exponential smoothing.

Sum of Count

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Multiplicative	None	Multiplicative	9	7	0.33	69.2%	147	0.500	0.000	0.000

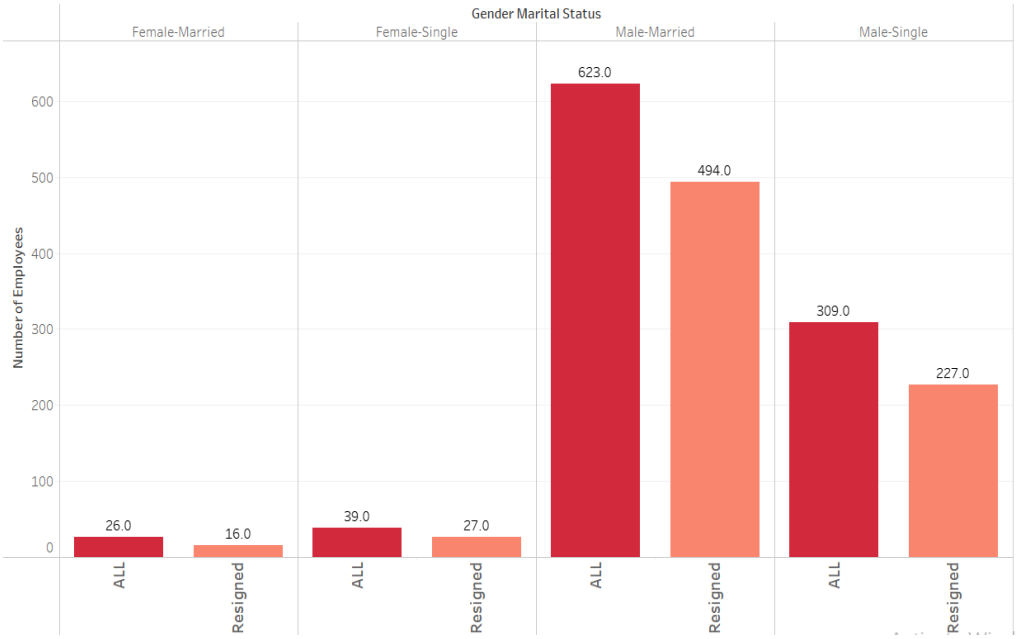
Insight 2 - Insights based on Gender and Marital Status



- It can be determined that Male employees tend to receive more net salary than Female employees, also Men are more likely to work longer Over time hours.

- Also we can see Married men are more likely to receive higher salaries and to work longer hours But,

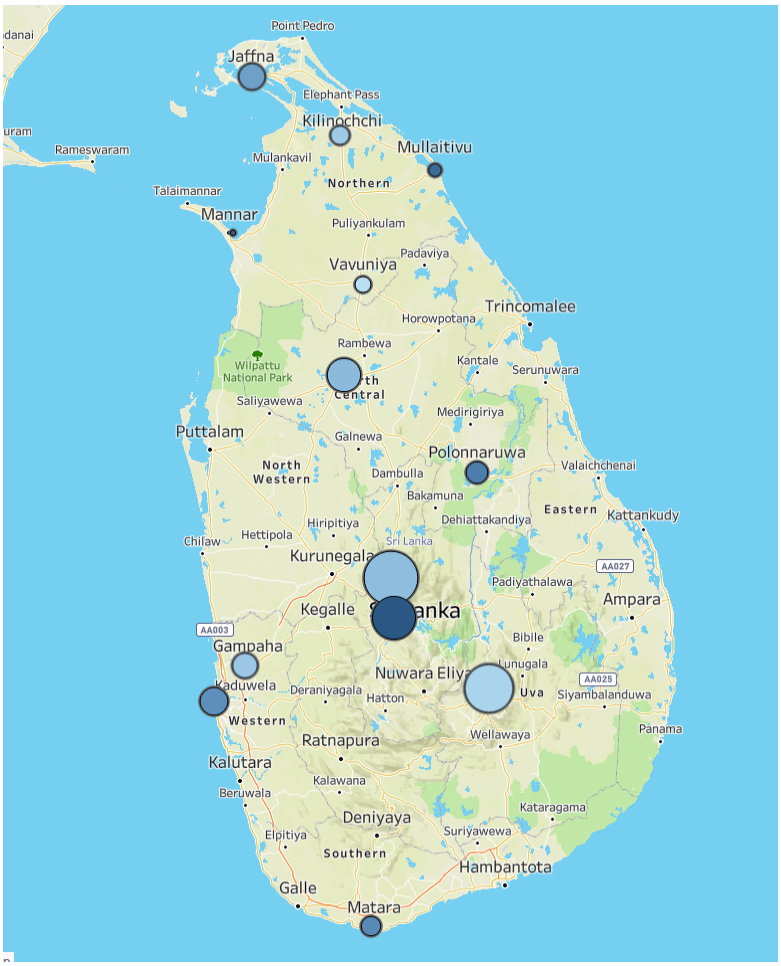
When it comes to women it's the other way around; Single women receive higher salaries and work longer hours.

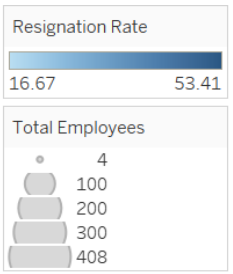


- We can see Male employees are the most resigned among all.

Male-Married resigned percentage is 79% while Male-Single percentage is 73%. In the first graph we could see when it comes to Male employees, the difference between Net Salary and OT hours are larger than Female Employees. Hence working for longer hours might be a reason for having a higher resignation rate.

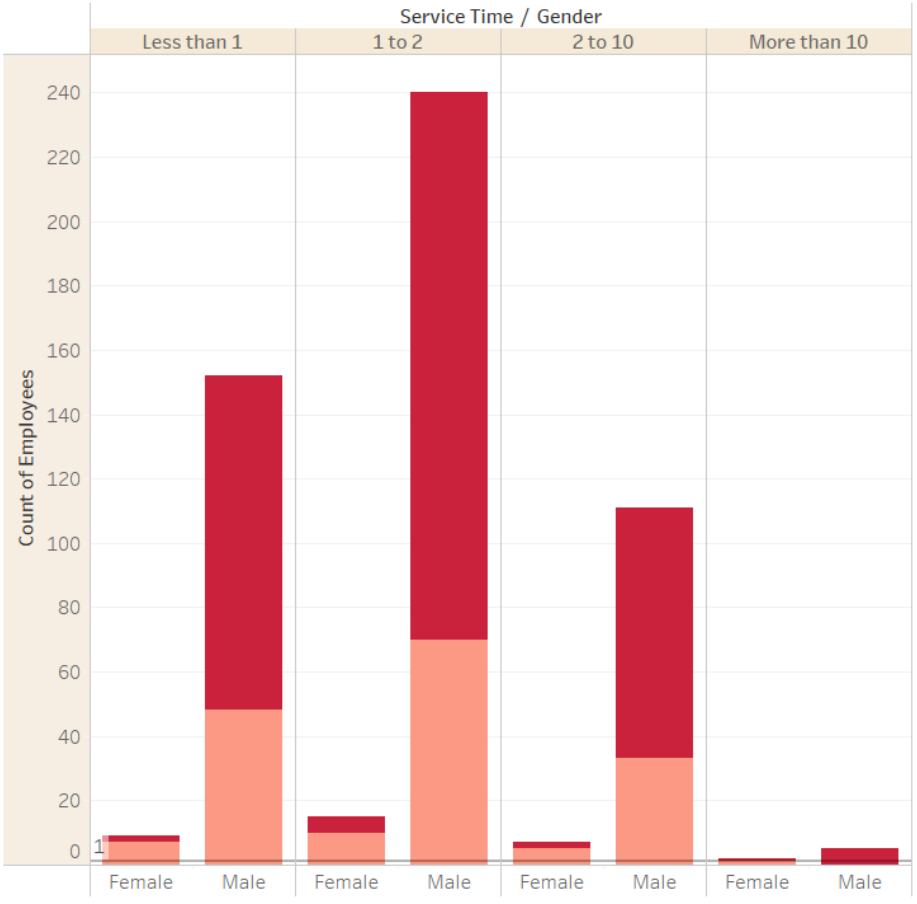
Insight 3 - District wise Employees and Sites measures

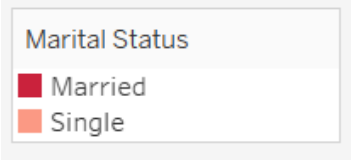




- According to the map in certain cities such as Mulativ, Polonnaruwa and Kandy Employee resignation rate is very high.
- Apart from **Kandy**, other cities with a large number of Employees working in, the rate of Resignation is low. Ex: Badulla, Mathale, Anuradhapura.
- Also in the central area more Employees are working, relative to the coastal side.

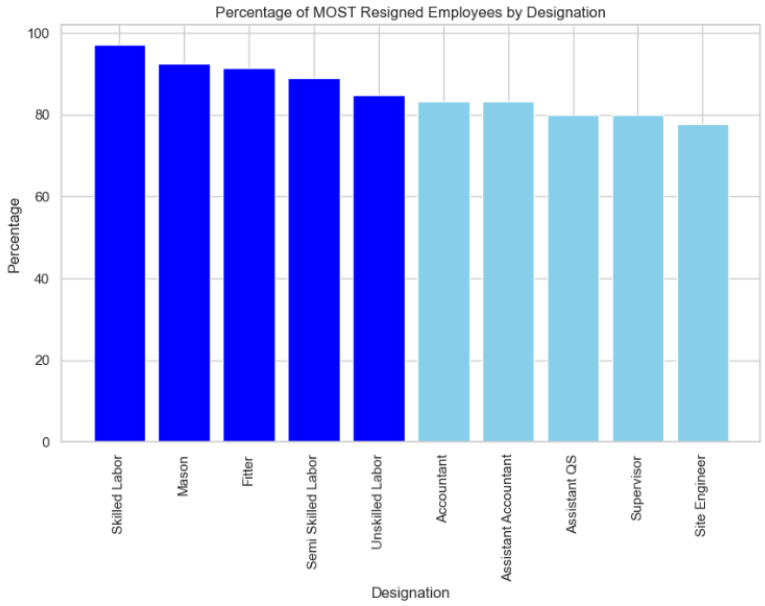
Insight 4 - Service Time upon resignation

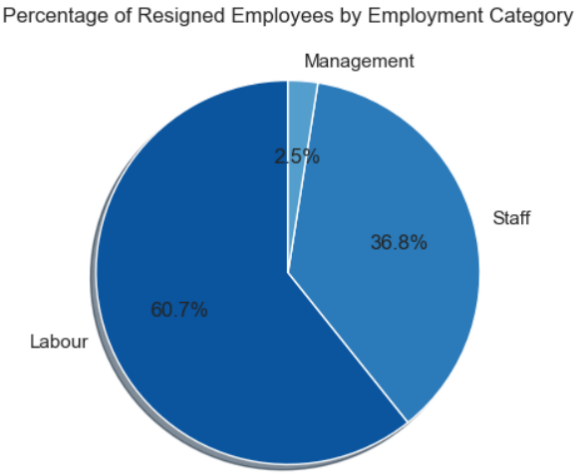




- According to this barplot, we can determine that most of the resignations were reported below two years of the service time.
- Hence, the company may need to implement some strategies such as providing continuous training and development opportunities and improving onboarding and orientation processes to keep the newcomers engaged and motivated.

Insight 5 - Resigned Employee by the designation and Employment Category





- According to the data, we can see the Labour department has the highest number of resigned employees. Also top 5 most resigned Designations are also Labour-Related. Hence prioritising the problems in the Labour department would be beneficial to reduce the attrition of employees.