CS3121 Introduction to Data Science

Employee Attrition in Marvelous Construction

Final Report

210518H Ranasinghe K.S
210353V Madhusanka G.I.D.L
210483T Prabashwara D.G.H
210588U Senarathna L.P.S.U.K.
210460V Perera I.T.M.

Group 18
9th May, 2024

Problem overview

We are given a dataset addressing employee attrition in Marvelous Construction. The goal is to analyze the dataset and find the possible causes for these resignations so that the CEO of Marvelous Construction can take the necessary actions to enhance employee retention in the company.

Dataset description

The provided dataset contains 6 csv files extracted from the ERP of Marvelous Construction. Details about each data file are as follows. Out of these datasets, holidays have no real use here. Similarly salary_dictionary_data is also there to understand the salary calculations from the different metrics given in the salary dataset. The other 4 datasets will be used in the analysis and hypothesis testing done next. Also we will be using the preprocessed employee dataset which we completed as a part of the previous submission for analysis.

File Name	Dimension	Remarks
employee	(997, 19)	This contains the target feature 'Status'
leaves	(1018, 6)	New employees have fewer leaves, while old employees have a higher number of leaves.
salary	(9035, 109)	Monthly addition/deduction breakdown is included
attendance	(224057, 10)	Late minutes = in time - shift start time
holidays	(120, 1)	-
salary_dictionary_data	(53, 4)	Contains information about the salary dataset.

Data pre-processing

Data preprocessing was performed in order to clean the data given, for a better analysis and gain insights.

Under data preprocessing, the employee data was analyzed. It was observed that there are missing values in “Year of Birth” and “Marital Status”. Thus it was decided to impute the missing data with a suitable value. When observing the available data in the “Year of Birth” column, it was noted that the data distribution was negatively skewed. It was decided to impute the missing values with the median. “Marital Status” contained categorical nominal data with NaN values. The possible values for this column were “Single” and “Married”. Since it is nominal and had only two possible values, the mode was used to impute the missing data with. If the class imbalance is high, using mode would have given less errors.

Next, the dataset was inspected for possible data quality issues. It was observed that there were mismatches between the gender (male/female) and the title (Mr. / Mrs. / Miss) present in the data, so it was resolved by manually correcting the mismatches. Since there were two columns indicating the designation (Designation ID, Designation) and religion (Religion ID, Religion) each, the ID columns were dropped as the ID and value columns uniquely map to each other. Also, since the “Status” (Active / Inactive) and “Inactive Date” were observed to match each other (null values in “Inactive Date” matches with the status being “Active” and an inactive date was available for each record with the status “Inactive”), the status column was dropped.

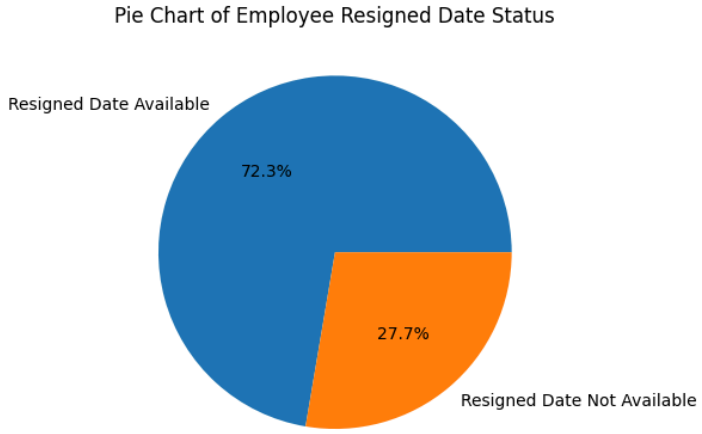
The next stage was data transformation. Under this process, firstly, all other missing data in the dataset was converted to NaN form as different columns/ features had represented missing values in their own unique ways. Secondly, all the date values were converted into timestamps. Thirdly, a conversion of the data types of the columns were done.

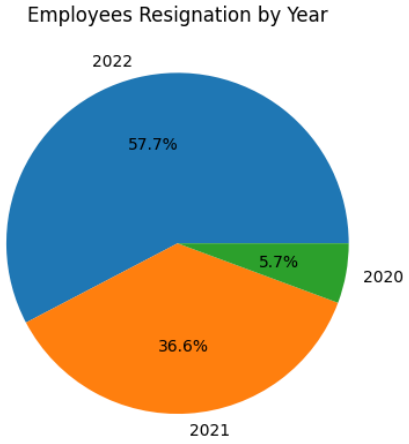
During the data preprocessing process, other data sets were also analyzed to gain possible insights in handling the main employee data set.

Insights from data analysis

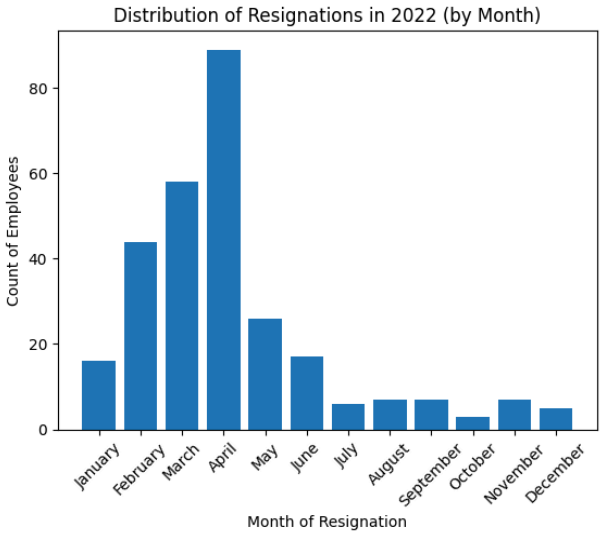
Insight 1

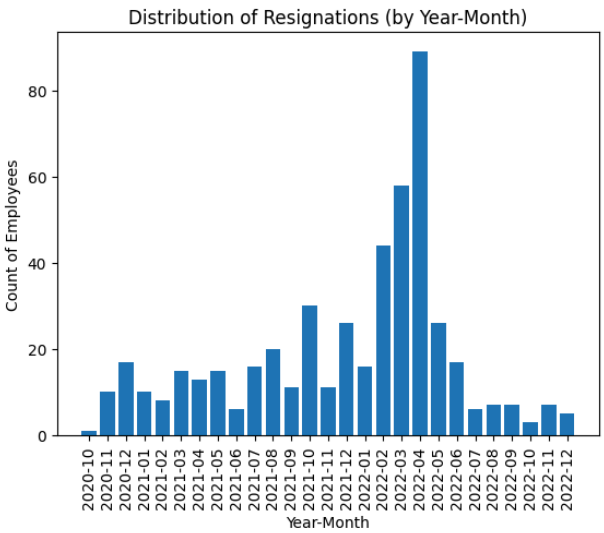
When analyzing the data from employee.csv file, it was observed that most of the inactive employees (employees having the value of 'Status' column as 'Inactive') have officially resigned from the company (as they have a valid date in the 'Date_Resigned' column).





When observing the resignation year (obtained by taking the year from the values in the 'Date_Resigned' column) of the employees who resigned, it was observed that the most have resigned in the year 2022, with a percentage of 57.7%.



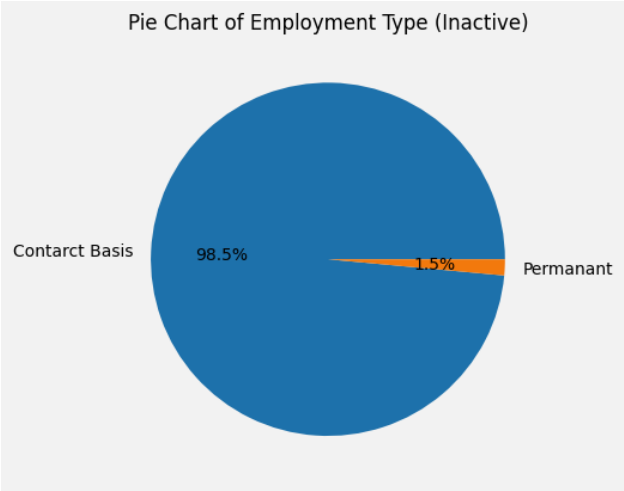


When taking the employees who have resigned in 2022, it can be observed that most employees have resigned in the first quarter. It is also noticeable that the resignations have increased month

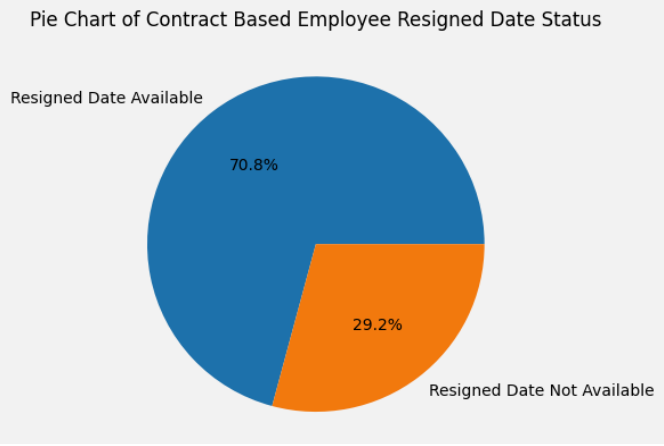
by month within the quarter, culminating in the month of April 2022 with the highest number of resignations. When analyzing the resignation month of the employees, it is evident that the highest number of employees have resigned during the period February-April 2022.

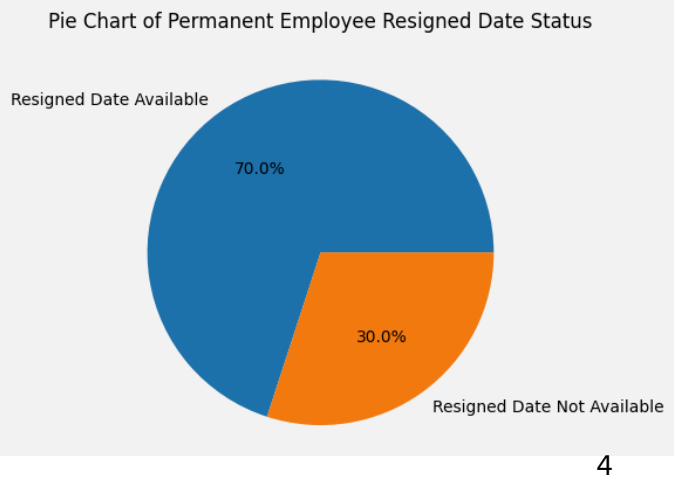
Insight 2

In the same file mentioned above, when observing the employment type of the inactive employees, it was noticed that the major sector was employed under contract. This might lead to resignations as there is less job security from the side of the employee as the offering of permanent employment is so few.



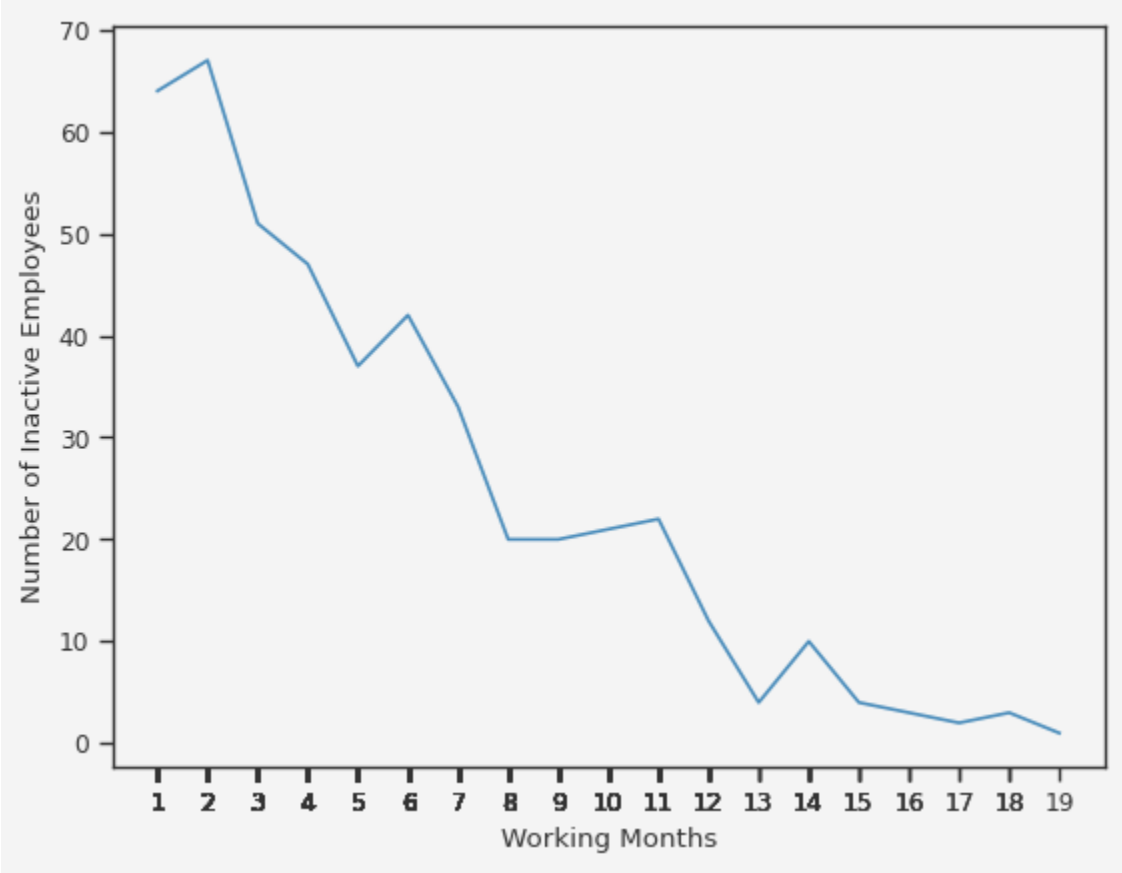
It also can be observed that employees leaving the company by resigning or without notice is not much dependent on the employment type as similar percentages of resignations are seen in both categories. This data was obtained by separating inactive employees under the employment type (‘Contract Basis’ and ‘Permanent’) and checking whether they have a valid date in the ‘Date_Resigned’ column under each employment type.





Insight 3

When observing the employment periods of the resigned employees we can see a trend of newcomers resigning more frequently. This was obtained using the salary dataset which contained the salary payments for each employee every month. Using aggregation we were able to obtain the no. of working months for each employee through which we obtained this graph.



We can see as much as 60 employees resigning within the 1st, 2nd and 3rd months of employment. While the longer term trend seems to be positive with fewer employee resignations. This trend in particular is unacceptable for the business. Recognizing this, it becomes imperative to prioritize the satisfaction and fulfillment of new employees' needs in order to enhance retention rates during their initial employment period.

Results of Hypothesis Testing

As this dataset focuses on employee attrition we chose hypotheses with regard to employee resignings. For that we chose the feature 'Status' in the employees.csv. We will be taking a significance level of 0.05 (5%) for all the tests.

1. Gender vs Status

Both these features are categorical with two unique values for both features. Therefore we chose the chi-square test for this hypothesis.

- H_0 : Gender does not significantly impact Employment Status.
- H_A : Gender significantly impacts Employment Status.
- **Test Results** :
    - Pearson Chi-square (df =2) = 3.4762
    - P-value = 0.1758

- **Conclusion** : Here p-value is higher than the significance level. Therefore the **null hypothesis is not rejected**. This means that there is no significance of Gender on Status. Which means Status is independent of Gender.

2. Average Earnings vs Status

Average earnings is a continuous variable. Point biserial test is used for comparisons between a binary variable and a continuous variable. So we chose the Point biserial test.

- H_0 : Average earnings do not significantly impact Employment Status.
- H_A : Average earnings significantly impacts Employment Status.
- **Test Results** :
    - Point biserial correlation = - 0.1915
    - P-value = $2.2881 * 10^{-7}$

- **Conclusion** : Here p-value is lower than the significance level. Therefore the **null hypothesis is rejected**. This means that there is a significance of Average Earnings on Status. Which means Employment Status is dependent on Average Earnings.

3. Average daily working hours vs Status

Average daily working hours is a continuous variable. Point biserial test is used for comparisons between a binary variable and a continuous variable. So we chose the Point biserial test.

- H_0 : Average daily working hours does not significantly impact Employment Status.
- H_A : Average daily working hours significantly impacts Employment Status.
- **Test Results** :
    - Point biserial correlation = - 0.0092
    - P-value = 0.8013

- **Conclusion** : Here p-value is higher than the significance level. Therefore the **null hypothesis is not rejected**. This means that there is no significance of average daily working hours on Status. Which means Status is independent of average daily working hours.

4. Leaves Count vs Status

Leaves count is a discrete ordinal variable while status is a categorical binary variable. So we chose the Spearman test.

- H_0 : Leaves count does not significantly impact Employment Status.
- H_A : Leaves count significantly impacts Employment Status.
- **Test Results** :
    - Spearman correlation = - 0.2242
    - P-value = $7.9014 * 10^{-13}$

- **Conclusion** : Here p-value is lower than the significance level. Therefore the **null hypothesis is rejected**. This means that there is a significance of leaves count on Status. Which means Employment Status is dependent on Leaves count.

5. Joined Date vs Status

Joined Date was converted to a timestamp during pre-processing. So it is a discrete ordinal variable while status is a categorical binary variable. So we chose the Spearman test.

- H_0 : Joined date does not significantly impact Employment Status.
- H_A : Joined date significantly impacts Employment Status.
- **Test Results** :
    - Spearman correlation = - 0.1077
    - P-value = $6.5761 * 10^{-4}$

- **Conclusion** : Here p-value is lower than the significance level. Therefore the **null hypothesis is rejected**. This means that there is a significance of the joined date on Status. Which means Employment Status is dependent on joined date.