Employee Attrition Analysis Report

Marvelous Construction

CS3121 - Introduction to Data Science

Name : Thanikan S.

Index No. : 200635E

Date : 16-07-2023

Note: Underlined file names contain links to the corresponding file on my project's GitHub repository.

Problem overview

Marvelous Construction is a major construction firm in Sri Lanka facing a concerning increase in employee attrition across its 35 construction sites. The CEO of Marvelous Construction seeks insights from available data to understand the underlying causes and develop effective strategies for employee retention. To address this problem, I was hired as a data scientist to analyze the provided dataset and uncover insights that can guide strategic decisions for improving employee retention.

The objective of this analysis is to uncover patterns and key factors associated with employee attrition by exploring employee-related variables such as gender, marital status, job role, date joined, date resigned, and year of birth. In addition, attendance, leave, and salary information were examined to understand how they impact employee attrition. By identifying the root causes of employee attrition, Marvelous Construction can take proactive measures to improve employee satisfaction, create a positive work environment, and establish policies for long-term engagement and loyalty. The findings from this analysis will serve as the foundation for evidence-based recommendations that aim to reduce attrition and foster a sustainable and productive workforce. Through descriptive, exploratory, and predictive data analysis, I aim to identify patterns, trends, and key drivers associated with attrition within the organization. These insights will empower the CEO to make informed decisions and implement effective strategies to mitigate attrition and enhance overall employee retention.

This report outlines the preprocessing steps undertaken on the dataset and presents the insights derived from the various data analysis techniques to reveal significant findings related to employee attrition at Marvelous Construction.

Dataset description

I was provided with a dataset that contains different files providing information about employee details, attendance, leaves, and salary details extracted from the ERP of Marvelous Construction. More details about each data file in the dataset are given below. Note that I described only the attributes of the files used for analysis, and I skipped the description of salary_dictionary.csv which I didn't use in my analysis.

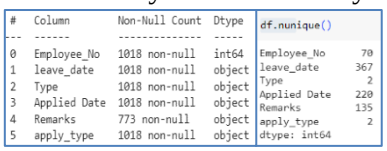
File Name	No. of Records	Attribute/Variable	Description	Data Type
employee.csv	997	Employee_No	Employee Identification Number	Nominal
		Title	Title using before name by an Employee (Mr or Miss or Ms)	Nominal
		Gender	Male or Female	Nominal
		Marital_Status	Single or married	Nominal
		Date_Joined	The date on which joined the company	Ordinal
		Date_Resigned, Inactive_Date	Date of Resignation	Ordinal
		Status	Target variable of the analysis shows whether the Employee is active or inactive	Nominal
		Employment_Category	Staff or Management or Labour	Nominal
		Employment_Type	Permanent or Contract Basis	Nominal

		Religion	Buddhist or Catholic or Hindu or Muslim	Nominal
		Designation	129 different designations (For example Driver, Account Clerk, Storekeeper)	Nominal
		Year_of_Birth	Birth year	Discrete
		Employee_Code, Name, Religion_ID, Designation_ID, Reporting_emp_1, Reporting_emp_2	Other attributes not used for further analysis	Nominal
leaves.csv	1018	Employee_No	Employee Identification Number	Nominal
		leave_date	The date on which the employee took a leave	Ordinal
		Type	Half-Day or Full-Day Leave	Nominal
		Applied Date	The date on which the employee applied for leave	Ordinal
		apply_type	Annual or Casual leave	Nominal
		Remarks	Another attribute with 135 unique records not used for further analysis	Nominal
salary.csv	9035	Employee_No	Employee Identification Number	Nominal
		Total Earnings_0, Total Earnings_2	Final salary in that month	Continuous
		Total Fixed	Fixed salary of the employee	Continuous
		<<105 Other Attributes>>	Other attributes not used for further analysis	Ordinal: month Nominal: Area and SiteNo Discrete: year Continuous: others
attendance.csv	224057	date	The date on which the employee started the shift	Ordinal
		out_date	The date on which the employee finished the shift	Ordinal
		Employee_No	Employee Identification Number	Nominal
		in_time	The time at which the employee started the shift	Continuous
		out_time	The time at which the employee finished the shift	Continuous
		Hourly_Time	Total worked time in hours	Continuous
		Shift_Start	Time at which shift starts	Continuous
		Shift_End	Time at which shift ends	Continuous
		id, project_code	Other attributes not used for further analysis	Nominal
holidays.csv	121	Unnamed single column	Holiday dates	Ordinal

Data pre-processing

➢ Preprocessing of leaves.csv

For the preprocessing of this file, Additionally, I used holidays.csv to check which days are the holidays. I carried out preprocessing as follows: (You can view my whole preprocessing task of leaves.csv in DS_Proj_leaves_preprocess.ipynb.



```
#  Column        Non-Null Count  Dtype
---  ------        --------------  -----
0  Employee_No   1018 non-null   int64
1  leave_date    1018 non-null   object
2  Type          1018 non-null   object
3  Applied Date  1018 non-null   object
4  Remarks       773 non-null    object
5  apply_type    1018 non-null   object
```

```
df.nunique()

Employee_No     70
leave_date     367
Type             2
Applied Date   220
Remarks        135
apply_type       2
dtype: int64
```

- Drop any duplicate records at first.
- Convert the columns leave_date and Applied Date to DateTime format to get all the dates in the same format.
- There were 466 records of annual type leaves, and 376 records of Half-day type leaves out of 1018 records. So, Type and apply_type may provide significant information for further analysis.
- Check for any significant separation between the informed and uninformed leaves. But it returned 974 records out of 1018 records were uninformed leaves. So, there is no significance with the Applied Date attribute.
- Remove the records in which leave was taken, but it was a holiday.
- Drop the columns which not needed for further processing (Applied Date and Remarks).
- Drop the duplicate records once again (There may be duplications by Applied Date and Remarks for the same leave_date).
- Some situations will produce implicit duplications as the employee taking Annual and Casual leave on the same day and the employee taking Full-day and Half-day type leave on the same date. So, remove both the situations by imputing such duplications with the mode of that attribute considering the records for that particular employee. And if only duplicate records are present for that employee, impute them with the mode of that attribute considering all the employee records.
- Group records by Employee_No and add Type_Full Day, Type_Half Day, apply_type_Casual, and apply_type_Anual attributes, which are the counts of leaves on that category by an employee.

➢ Preprocessing of attendance.csv

I carried out preprocessing as follows: (You can view my whole preprocessing task of attendance.csv in DS_Proj_attendance_preprocess.ipynb)

- Impute missing values of hourly time by finding the difference between out_time and in_time.
- Replace strings starting with "24" to "00" in the out_time column since it will raise an error in parsing out_time from string format to time format.
- Drop any duplicate records.
- Parse the in_time and Shift_Start columns to time format, and then calculate the Late_minutes by subtracting shift_time from in_time. Finally, put zero for the non-late records.
- Drop columns not needed for further processing, such as id, project_code, date, out_date, out_time, in_time, Shift_Start, and Shift_End.
- Group records by 'Employee_No' and calculate the average of Hourly_Time and Late_minutes.

➢ Preprocessing of salary.csv

I carried out preprocessing as follows: (You can view my whole preprocessing task of salary.csv in DS_Proj_salary_preprocess.ipynb)

- Drop the duplicate records and check for whether both Total Earnings_0 and Total Earnings_2 have nonzero values. It returned nothing as the situation. So, I added both Earnings.
- Since many redundant columns which not needed for further processing, select Employee_No, Earnings, and Total Fixed and drop other columns.
- Group records by 'Employee_No' and calculate the average for the Earnings and Total Fixed.

➢ Preprocessing of employee.csv

The major preprocessing task was carried out in this file. I carried out preprocessing as follows: (You can view my whole preprocessing task of employee.csv in DS_Proj_employee_preprocess.ipynb)

- Drop any duplicate records.
- Replace "0000" with null in 'Year_of_Birth'
- There were 223 nulls in Date_Resigned and there was nothing null in Inactive_Date when the status was Inactive. Also, only 6 records were found as having different values for Date_Resigned and Inactive_Date when both are not null. So, imputing Date_Resigned with the same date as Inactive_Date is the best choice for imputing null values of Date_Resigned when Status is Inactive.

- Replace all the cells with '\\N', "0000" and '0000-00-00' with null value.
- Checked for incompatible Gender and Title. I found some of them, and I gave priority to Gender. I filled Title as Ms for Female and Mr for Male as they are used for addressing a woman and man without specifying their marital status.
- Parse columns with dates to datetime format to get all dates in the same format.
- I checked for non-null values and data types. Since Reporting_emp_1 and Reporting_emp_2 is with too much of missing values, I dropped those two columns. Also, I dropped Employee_Code, Name, Religion_ID, Designation_ID, and Inactive_Date as they are not needed for further Processing.
- Merge the preprocessed files of leaves.csv, attendance.csv, and salary.csv to the employee.csv by considering the common attribute Employee_No. Then check for the non-null values and data types.
- Encode all the categorical nominal variables using Label Encoding.
- Replace null values with a specific value -1, to represent missing values of encoded Marital_Status.

Figure: Before Data Integration

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 997 entries, 0 to 996
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Employee_No           997 non-null    int64
 1   Employee_Code         997 non-null    int64
 2   Name                  997 non-null    object
 3   Title                 997 non-null    object
 4   Gender                997 non-null    object
 5   Religion_ID           997 non-null    int64
 6   Marital_Status        930 non-null    object
 7   Designation_ID        997 non-null    int64
 8   Date_Joined           997 non-null    datetime64[ns]
 9   Date_Resigned         764 non-null    datetime64[ns]
 10  Status                997 non-null    object
 11  Inactive_Date         764 non-null    datetime64[ns]
 12  Reporting_emp_1       61 non-null     object
 13  Reporting_emp_2       0 non-null      float64
 14  Employment_Category   997 non-null    object
 15  Employment_Type       997 non-null    object
 16  Religion              997 non-null    object
 17  Designation           997 non-null    object
 18  Year_of_Birth         894 non-null    float64
dtypes: datetime64[ns](3), float64(2), int64(4), object(10)
memory usage: 148.1+ KB
```

Figure: After Data Integration

```
#   Column                Non-Null Count  Dtype
---  ------                --------------  -----
0   Employee_No           997 non-null    int64
1   Title                 997 non-null    object
2   Gender                997 non-null    object
3   Marital_Status        930 non-null    object
4   Date_Joined           997 non-null    datetime64[ns]
5   Date_Resigned         764 non-null    datetime64[ns]
6   Status                997 non-null    object
7   Employment_Category   997 non-null    object
8   Employment_Type       997 non-null    object
9   Religion              997 non-null    object
10  Designation           997 non-null    object
11  Year_of_Birth         894 non-null    float64
12  Hourly_Time           742 non-null    float64
13  Late_minutes          742 non-null    float64
14  Type_Full Day         57 non-null     float64
15  Type_Half Day         57 non-null     float64
16  apply_type_Anual      57 non-null     float64
17  apply_type_Casual     57 non-null     float64
18  Earnings              719 non-null    float64
19  Total Fixed           719 non-null    float64
```

- **Impute missing values for Year_of_Birth and Marital_Status**
  - The following table shows the association between variables found by different tests based on both variable types:

Variable 1	Variable 2	Test used and the statistic from the test				Selected Features for Variable 1
		Spearman's	Point-biserial	Chi-squared	p-value	
Marital_Status	Title			2.99430e+01	3.14741e-07	Chi-squared test and Point-biserial correlation coefficient were carried out to quantify the association of variables with Marital_Status. So, I compared the p-values to select the best features from available features for the model to predict missing values of Marital Status. I select features where p-value < 0.001(e-03). Following are such features that are highly correlated with Marital_Status: - Title - Gender - Employment_Category - Designation - Year_of_Birth
	Gender			2.46240e+01	6.96780e-07	
	Status			2.65471e+00	1.03243e-01	
	Employment_Category			4.44030e+01	2.28042e-10	
	Employment_Type			4.23900e+00	3.95055e-02	
	Religion			1.53647e+01	1.53009e-03	
	Designation			2.53787e+02	2.81582e-11	
	Year_Joined		-0.05964		6.90762e-02	
	Year_of_Birth		-0.60940		3.88341e-86	
	Hourly_Time		0.02778		4.68943e-01	
	Late_minutes		-0.07599		4.72745e-02	
	Earnings		0.01698		6.61985e-01	
	Total Fixed		-0.04841		2.12478e-01	
	Type_Full Day		-0.01885		8.94453e-01	
	Type_Half Day		-0.12862		3.63482e-01	
	apply_type_Anual		-0.03722		7.93324e-01	
	apply_type_Casual		-0.08277		5.59646e-01	
Year_of_Birth	Title		0.09324		5.26925e-03	Spearman's correlation coefficient and Point-biserial correlation coefficient were carried out to quantify the association of variables with Year_of_Birth. So, I compared the p-values to select the best features from available features for the model to predict missing values of Year_of_Birth. I select features where p-value < 0.001(e-03). Following are such features that are highly correlated with Year_of_Birth: - Gender - Employment_Category - Marital_Status
	Gender		-0.11888		3.67628e-04	
	Status		-0.03008		3.69005e-01	
	Employment_Category		0.16231		1.06866e-06	
	Employment_Type		-0.09317		5.30601e-03	
	Religion		0.07289		2.93101e-02	
	Designation		-0.05200		1.20290e-01	
	Year_Joined	0.10727			1.31702e-03	
	Marital_Status		-0.60940		3.88341e-86	
	Hourly_Time	-0.06548			8.93662e-02	
	Late_minutes	0.12634			1.01208e-03	
	Earnings	0.03224			4.10474e-01	
	Total Fixed	0.10826			5.57938e-03	
	Type_Full Day	-0.02243			8.69638e-01	
	Type_Half Day	0.05101			7.08904e-01	
	apply_type_Anual	-0.00487			9.71566e-01	
	apply_type_Casual	-0.04161			7.60734e-01	

  - Year_of_Birth and Marital_Status are more highly correlated than any other pair according to the above results. So, I checked how many records are missing both the above two attributes, and it returned 9 records. Hence, I planned to first impute the records with missing only one attribute by prediction from a model, and then fill the remaining with suitable central tendency measures.
  - Creating models and using them for prediction.
    - Extract records without any missing values in both Year_of_Birth and Marital_Status (836 records were found).
    - Split the records into train and test sets.
    - Using sklearn's GridSearchCV tune the parameters of the RandomForest model by training with train test and testing with test set from the above step.
    - Using the best parameters gained from the above step create the best models for each Marital_Status and Year_of_Birth target variables.

```
RandomForestRegressor(min_samples_split=15, n_estimators=300, random_state=10)
```

For predicting Year_of_Birth (R-squared score of this model: 0.466)

```
RandomForestClassifier(min_samples_split=20, random_state=10)
```

For predicting Marital_Status (Accuracy of this model:0.893)

    - Now train the model with both train and test. Then predict and impute the missing values.
  - Impute the remaining 9 records where both variables missing with a suitable central tendency measure. I used mode to fill Marital_Status and median to fill Year_of_Birth (I used median instead of mean as it skewed and contain outliers).

- Transfer all the filled null values to the main data frame.
- Check for any other null in the data frame. It returned that 'Date_Resigned', 'Hourly_Time', 'Late_minutes', 'Type_Full Day', 'Type_Half Day', 'apply_type_Anual', 'apply_type_Casual', 'Earnings', 'Total Fixed' are now having null values. It's trivial that Date_Resigned can be null for Inactive Employees. But we don't have any idea about other variables. For example, if we get information on Employees whose records not in leaves.csv didn't take any leave, then we must put zero for null values of leave counts and it may have a significant association with the target variable Status. Hence, I skipped the imputation of null values for the above variables.
- Calculate the Resigned_Age by finding the duration between Year_of_Birth and Date_Resigned.
- Calculate the Worked_Duration by finding the duration between Date_Joined and Date_Resigned and fill the remaining null with the current age of employees.

Insights from data analysis

After the preprocessing, I used employee_preprocess_200635E.csv with Python seaborn and matplotlib libraries for the visualizations. Also, I used the Power BI tool, and I derived the following insights: (Link to Power BI file: DataAnalysis.pbix, Link to Data Analysis python notebook: DS_Proj_Data_Analysis.ipynb)

1) The following are the quantity measure of association of Status with other variables:

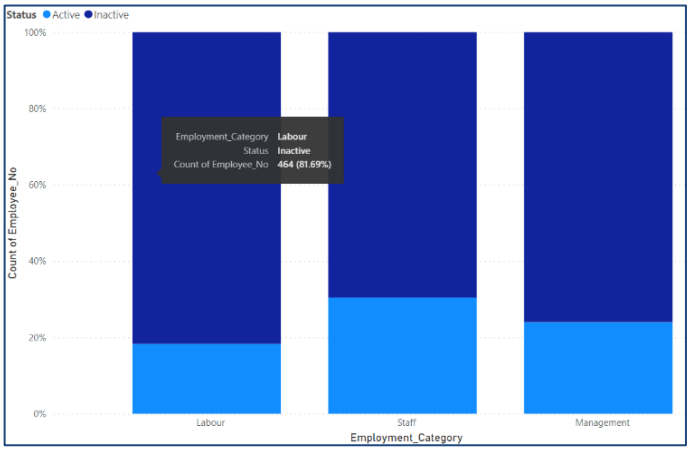
```
Chi-square test between Title and Status
Chi-square test statistic: 4.26213448068514
p-value: 0.11871053355209335
Chi-square test between Gender and Status
Chi-square test statistic: 3.658364238835724
p-value: 0.05578837958109029
Chi-square test between Marital_Status and Status
Chi-square test statistic: 3.4889855592529972
p-value: 0.06177843105135954
Chi-square test between Employment_Category and Status
Chi-square test statistic: 19.420488818889872
p-value: 6.065888667228195e-05
Chi-square test between Employment_Type and Status
Chi-square test statistic: 6.635632103042585
p-value: 0.00999587204002437
Chi-square test between Religion and Status
Chi-square test statistic: 4.0461987006202955
p-value: 0.256518523656958
Chi-square test between Designation and Status
Chi-square test statistic: 232.601623555578
p-value: 4.5254947645007056e-08
Chi-square test between Marital_Status and Status
Chi-square test statistic: 3.4889855592529972
p-value: 0.06177843105135954
```

```
Point-biserial correlation coefficient between Year_of_Birth and Status:
correlation coefficient: 0.03044778405416198
p-value: 0.33684652520741376
Point-biserial correlation coefficient between Year_Joined and Status:
correlation coefficient: -0.07934992199581392
p-value: 0.012199936211714233
Point-biserial correlation coefficient between Worked_Duration and Status:
correlation coefficient: 0.32161001278491747
p-value: 2.003597546235038e-25
Point-biserial correlation coefficient between Resigned_Age and Status:
correlation coefficient: 0.03293762143847391
p-value: 0.2988081179913326
Point-biserial correlation coefficient between Hourly_Time and Status:
correlation coefficient: 0.008797931359329656
p-value: 0.8109090996148427
Point-biserial correlation coefficient between Late_minutes and Status:
correlation coefficient: 0.02234419736193237
p-value: 0.5433874976436639
Point-biserial correlation coefficient between Type_Full Day and Status:
correlation coefficient: 0.30087297689488
p-value: 0.022954448932351584
Point-biserial correlation coefficient between Type_Half Day and Status:
correlation coefficient: 0.3025643469146815
p-value: 0.02216028713844025
Point-biserial correlation coefficient between apply_type_Anual and Status:
correlation coefficient: 0.2637032468449057
p-value: 0.047474296855775484
Point-biserial correlation coefficient between apply_type_Casual and Status:
correlation coefficient: 0.2921269650637013
p-value: 0.027454765346372954
Point-biserial correlation coefficient between Earnings and Status:
correlation coefficient: 0.19793774646363876
p-value: 8.73071137436596e-08
Point-biserial correlation coefficient between Total Fixed and Status:
correlation coefficient: 0.22304142666464669
p-value: 1.4793732398589018e-09
```

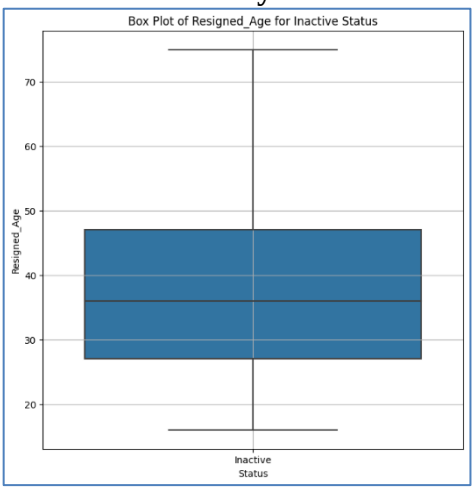
From the above results, we can conclude that **Employment_Category, Designation, Worked_Duration, Earnings, and Total Fixed are highly correlated with the Status.** I used the p-value of each test to derive the above result by selecting attributes where the p-value < 0.001(e-03).

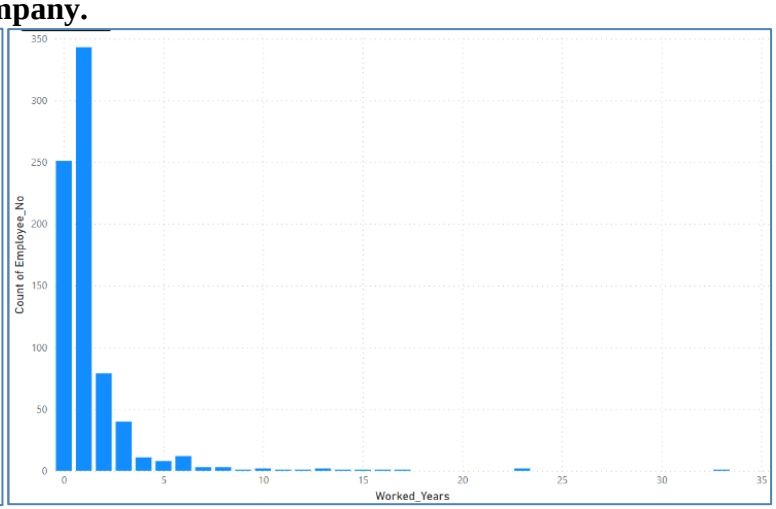
2) **The highest percentage of Employee attrition happened in Labour Category and the lowest was in Staff Category.**

I created a 100% stacked column chart to show the percentage share between Active and Inactive employees among Employment_Category to visualize how this variable influence on Status.



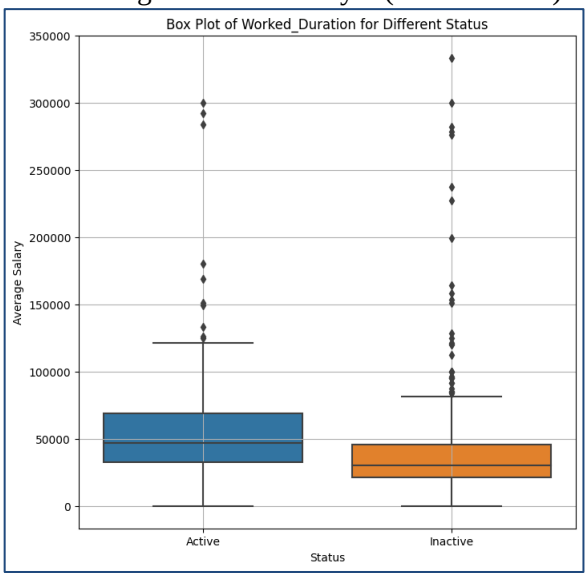
3) **About 50% of the employees resigned from their job between the ages 25 and 45 with a median age of nearly 35, and most of the resignations happened before work for at least 3 years in the company.**

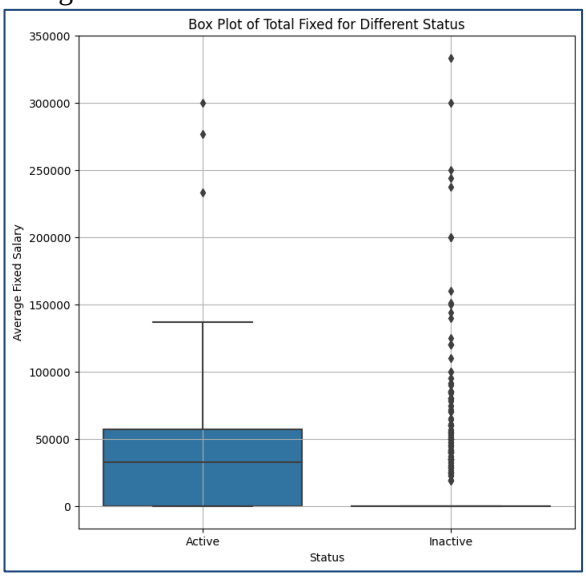




I used a boxplot to show the distribution of Resigned_Age among Inactive employees, and I visualize how count of employee who are resigned varies with Worked_Year using Bar chart.

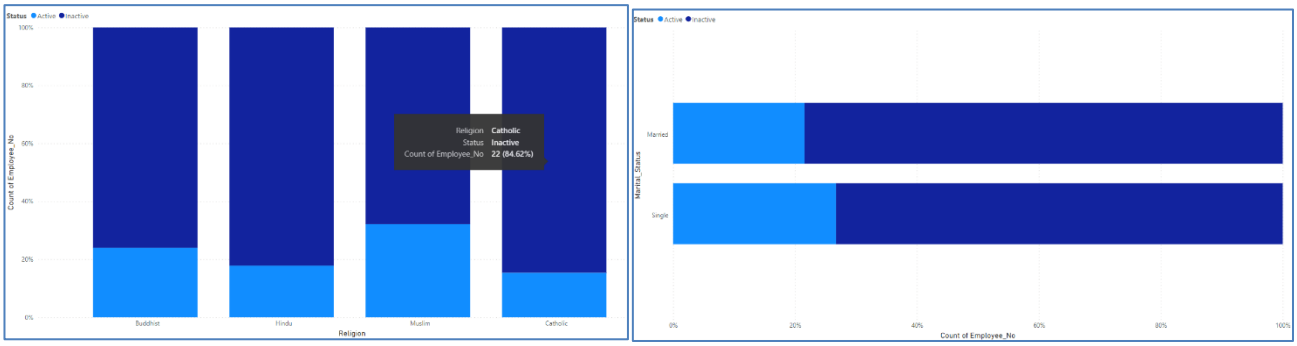
4) I created the following two boxplots which show the distribution of average salary (Earnings) and average fixed salary (Total Fixed) among Active and Inactive Statuses.





From the above two visualizations, we can observe that the average fixed salary distribution for Inactive employees is near zero and the average salary and the average fixed salary distribution is comparatively lower than Active employees. **Hence, providing a fixed salary for Employees and Increasing the Total Earnings of Employees will improve Employee retention.**

5) **The highest percentage of attrition was happed with Hindus and Catholics than Buddhists and Muslims, and Married Employees Resigned the job at Highest percentage than Singles.**



I created a 100% stacked column chart to show the percentage share between Active and Inactive employees among religions, and I created a 100% stacked bar chart to show the percentage share between Active and Inactive employees among Marital_Status.