CS3121 Introduction to Data Science

Employee Attrition in Marvelous Construction

Group M

M.A.S.N Aththanayake- 210055J
Pranavan Subendiran - 210491P
J.J Wickramasinghe- 210706H
M.A.C.L Mallikarachchi- 210363C
P.M.C. Nayanathara - 210417X

Problem overview

Marvelous Construction, a construction firm operating in Sri Lanka, is facing a high rate of employee resignations across its 35 sites. The company's Human Resources department seeks to understand the reasons behind this trend and implement strategies to reduce employee turnover.

Dataset description

File Name	# Records	Dataset descriptions
employee	631	You can find basic information related to employees through this csv file.
leaves	237	All information with regards to office leaves are mentioned in here. It specifies if a leave was half or full, annual or casual along with additional remarks.
salary	2632	Monthly addition/deduction breakdown is included along with Net salaries.
attendance	60354	Employee attendance related information is included. Shift beginning and ending times provide additional information related to their working hours.

Data pre-processing

1. **Employee table**

   Columns:
   Employee_No, Employee_Code, Name, Title, Gender, Religion_ID, Marital_Status, Designation_ID, Date_Joined, Date_Resigned, Status, Inactive_Date, Reporting_emp_1, Reporting_emp_2, Employment_Category, Employment_Type, Religion, Designation, Year_of_Birth

- We dropped the following columns as they seems to be less important: *Employee_Code, Name, Religion_ID, Designation_ID, Reporting_emp_1, Reporting_emp_2*
- Then we corrected the values of Title based on *Gender.*
- Then we corrected the values of *Date_Resigned, and Inactive_Date* using *Date_Resigned, Status, Inactive_Date.*
- We checked the distribution of *Year_of_Birth.* It was neither normal nor skewed. It was irregular. So, we planned to impute using K-Nearest Neighbours.
- We encoded the values appropriately and did KNN imputation. (Imputing *Marital_Status* was also done.)

	Employee_No	Title	Gender	Marital_Status	Date_Joined	Date_Resigned	Status	Inactive_Date	Employment_Category	Employment_Type	Religion	Designation	Year_of_Birth
0	347	Mr	Male	Married	12/8/1993	\N	Active	\N	Staff	Permanant	Buddhist	Driver	1965.0
1	348	Mr	Male	Married	3/14/1995	\N	Active	\N	Staff	Permanant	Buddhist	Driver	1973.0
2	349	Mr	Male	Married	1/27/1988	6/28/2021	Inactive	6/28/2021	Staff	Permanant	Buddhist	Account Clerk	1974.0
3	351	Ms	Female	Married	10/1/1999	1/31/2022	Inactive	1/31/2022	Staff	Permanant	Catholic	Purchasing Officer	1974.0
4	352	Mr	Male	Married	1/26/2001	\N	Active	\N	Staff	Permanant	Buddhist	Store Keeper	1980.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Employee_preprocessed: DS_Proj_EmployeeTable-15.ipynb

2. Attendance table

Columns:
id, project_code, date, out_date, employee_no, in_time, out_time, Hourly_Time, Shift_Start, Shift_End

We first filtered out the employees which are in the employee's table.
Each individual employee had many records and we concatenated and got the rows to an employee-wise format.

Then we came up with a set of new features by engineering the existing features:

Column	Description
1. Average work time	(out_time-in_time)/ no. of days
2. Average late hours	(In_time- shift_start_time)/ no. of days
3. Average leave early hours	(out_time- shift_end_time)/ no. of days
4. Project codes	Concatenated project codes according to employee
5. Absent count	Total count of days where In_time was equal to out_time

attendance_final

	Employee_No	Average_work_Time	Average_late_hours	Average_leave_early_hours	Project_Codes	absent count
0	347	8.72	0.18	-0.87	{1.0, 193.0}	7
1	348	11.81	-0.30	-3.83	{1.0, 193.0, 195.0, 194.1, 197.0, 198.1, 196.0...	135
2	349	8.91	0.46	-1.37	{1.0, 193.0}	5
3	351	8.37	-0.00	-0.34	{1.0}	1
4	352	10.54	0.50	-3.04	{187.0}	4
...	...	...	...	...	...	...
735	2836	8.17	0.05	-0.19	{1.0}	1
736	2890	11.13	0.05	-2.92	{1.0, 206.0}	2
737	2972	8.51	-0.20	-0.29	{1.0}	2
738	2973	10.57	-0.33	-2.00	{194.1, 196.0}	1
739	3041	9.69	-0.12	-1.07	{1.0}	4

740 rows × 6 columns

Attendance_final.df : attendance_preprocessed.ipynb

3. Salaries table

Columns: Employee_No, Amount, month, year, <<a total of 102 columns>>

This table had lot of redundant columns but useful columns were Total Earnings_2 , Net Salary and Total Deduction

Column	Description
Important Information with regard to preprocessing columns	Number of cases where Net Salary equals Total Earnings_2 minus Total Deduction: 7442 Net salaries with massive differences were dropped Net salaries with 0 were dropped Difference!=0 ->1000+ and the ones with a difference were mainly due to stamp charges.
	After noticing there were multiple entries for the same employee we decided to get the average Salary for each employee

	Employee_No	Average Salary
0	347	34529
1	348	61223
2	351	45604
3	352	31156
4	354	69234
...	...	...
641	2836	37175
642	2890	48141
643	2972	54481
644	2973	71108
645	3041	27841

	Employee_No	Total Earnings_2	Net Salary	Total Deduction
0	347	37412.43	35107.43	0.00
1	347	35356.81	33051.81	0.00
2	347	38409.95	0.00	0.00
3	347	36325.83	34020.83	0.00
4	347	37038.35	34733.35	2305.00
...	...	...	...	...
9030	3043	24310.00	24310.00	0.00
9031	3044	26010.00	25985.00	25.00
9032	3045	26100.00	26075.00	25.00
9033	3084	80000.00	12933.33	67066.67
9034	3095	0.00	0.00	0.00

9035 rows × 4 columns

Salaries_final.df: salary_preprocessing.ipynb

4. **Leaves table**

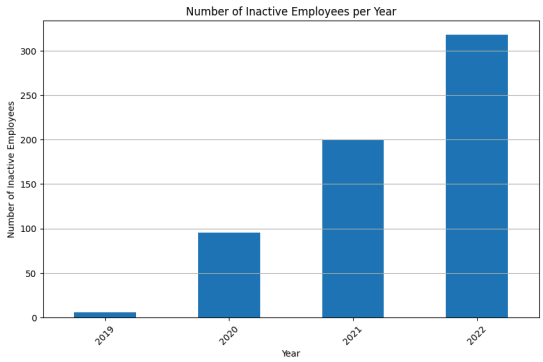
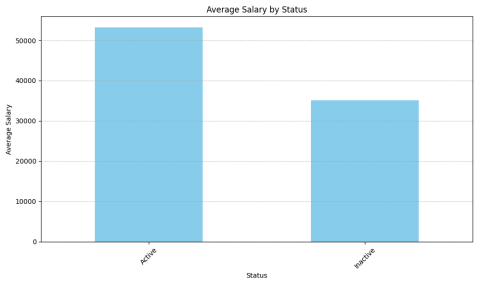
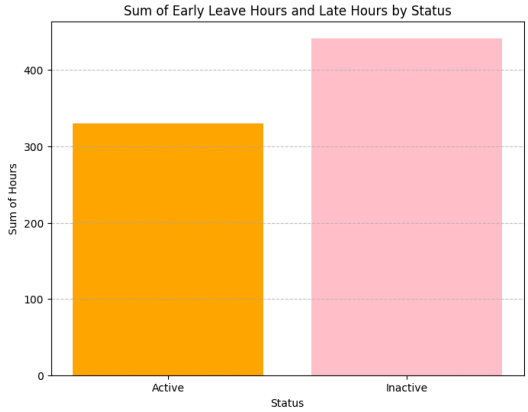
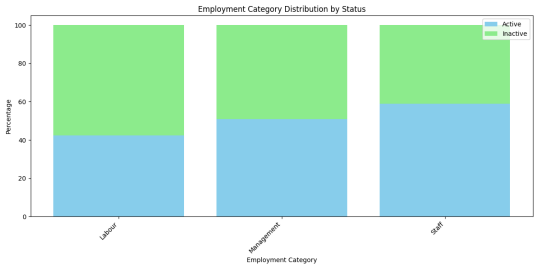
   Columns: Employee_No , leave_date, Type, Applied Date, Remarks, apply_type

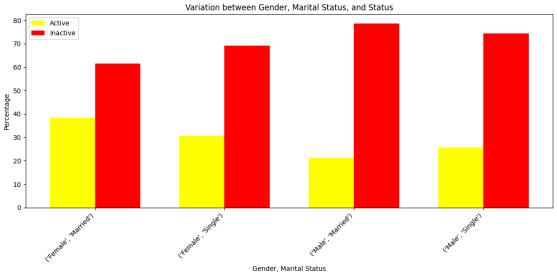
Column	Description
Preprocessed information	Leaves were concatenated according to employee_nos. Separation was done based on leave type (Half/Full) and apply type (Annual/Casual). Furthermore encoding was done as 1 if remarks were provided and 0 elsewise.

Employee_No	Half_Day_Count	Full_Day_Count	Anual_Count	Casual_Count
347	6	17	11	12
348	7	6	5	8
351	6	4	0	10
356	6	7	0	13
373	7	17	16	8
376	0	5	0	5
393	0	7	6	1
421	4	19	16	7
423	9	17	17	9
425	3	14	11	6

Leaves_final.df:
leaves_preprocessed.ipynb

Insights from data analysis

Insight 1		This bar graph shows that the number of employees leaving the organization has increased on an yearly basis exponentially.
Insight 2		It can be seen that employees with a classified status “Inactive” earn a relatively lower average salary as depicted by the bar chart.
Insight 3		This bar graph shows that the sum of early leave hours and late hours (now negated in the bar chart, the higher the value; the more time they’ve spent working) is higher for Inactive employees.
Insight 4		This stacked bar chart gives a clear understanding of active(blue) and inactive(green) employees based on the respective employment categories. For ease of understanding, the values are depicted as percentages out of the employment categories. The Labor category has the most inactive employees.

Insight 5		This double bar graph shows the distribution between genders and marital statuses against their active and inactive status as a percentage. Inactivity is especially high in men; out of which married men are higher.

Results of Hypothesis Testing

Hypothesis testing notebook link : DS_Proj_HypoTesting-15.ipynb

No.	Hypothesis	Conclusion
1.	H0: Average salary of employees doesn't significantly impact employee attrition.(independency) Ha: Average salary of employees of the shift impacts the employee attrition.(dependency)	F-statistic: 11.434787865157816 p-value: 9.42794173030303e-14 Hypothesis testing result: H0 is rejected at 5.0% significance level. H0 is rejected at 10.0% significance level. H0 is rejected at 20.0% significance level. **It seems there is a significant impact of salary in employee attrition.**
2.	H0: Duration of the shift doesn't significantly impact the employee attrition.(independency) Ha: Duration of the shift impacts the employee attrition.(dependency)	F-statistic: 2.8521412986754164 p-value: 0.00610116603960228 Hypothesis testing result: H0 is rejected at 5.0% significance level. H0 is rejected at 10.0% significance level. H0 is rejected at 20.0% significance level. **We can conclude that the duration of the shift significantly impacts employee attrition.**
3.	H0: Arriving late to the shift doesn't significantly impact employee attrition.(independency) Ha: Arriving late to the shift significantly impacts employee attrition.(dependency)	F-statistic: 3.337280910948583 p-value: 0.0009222669842211066 Hypothesis testing result: H0 is rejected at 5.0% significance level. H0 is rejected at 10.0% significance level. H0 is rejected at 20.0% significance level. **Arriving late to the shift significantly impacts employee attrition.**
4.	H0: Employment category doesn't significantly impact the employee attrition.(independency) Ha: Employment category significantly	Chi-Squared Test Results: Chi-Squared Statistic: 11.170402061107495 p-value: 0.0037529952272210075 Degrees of Freedom: 2

	impacts the employee attrition.(dependency)	Hypothesis testing result: H0 is rejected at 5.0% significance level. H0 is rejected at 10.0% significance level. H0 is rejected at 20.0% significance level. **There is a significant association between the employment category('Staff', 'Management', 'Labour') and employee attrition.**
5.	H0: Employment type doesn't significantly impact employee attrition.(independency) Ha: Employment type significantly impacts the employee attrition.(dependency)	Chi-Squared Test Results: Chi-Squared Statistic: 3.3261469638567682 p-value: 0.06818643848637372 Degrees of Freedom: 1 Hypothesis testing result: Failed to reject H0 at 5.0% significance level. H0 is rejected at 10.0% significance level. H0 is rejected at 20.0% significance level. **As H0 is rejected at 5 % significance there is no significant association between employment type('Permanent' or 'Contract Basis') and employee attrition.**