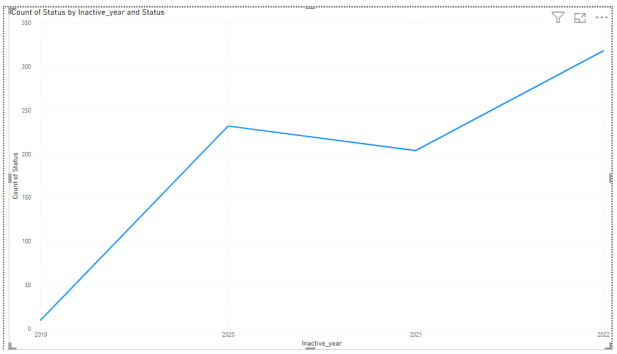
Final Report-200487B

Problem Overview

Client	Marvelous Construction
Issue	Increment in the number of employee resignations recently
Task	Giving insights on employee resignation on ● Descriptive ● Exploratory ● Predictive manner with could be helpful for preventing future employee resignations.
Targeted audience	CEO of Marvelous Construction

The following table describes what is intended to be archived in each phase of this analysis.

Descriptive	Identify trends, and describe the dataset
Exploratory	Identify possible reasoning behind employee resignation
Predictive	Identify employees who have a higher probability to resign in the future



graph_01

The problem can be clearly visualized by the graph_01 where the number of employee resignations is plotted against the resignation year. It is clearly observable that sudden increment of resignation.

Data set description

The data set consists of Employee, attendance, leaves, salary, and holiday datasets extracted from the Clients' ERP system.

Column description of impact features

Employee *(Size: 136 kB, Number of Attributes: 19, Number of Records: 997)*

- Employee_ID: The employee's identification number.
- Name: The employee's full name.
- Title: Mr/Mrs/Ms
- Gender: The employee's gender (male/female).

- Religion_ID: Identifier for the employee's religion.
- Marital_Status: Whether the employee is Married/Un-married
- Designation_ID: Identifier for the employee's designation.
- Designation: Employee's designation
- Date_Joined: Date of the employee joining the company
- Date_Resigned: Date of the employee resigning
- Status: Whether the employee is Active/Inactive
- Inactive_Date: Date when the employee is inactive from
- Reporting_emp_1/Reporting_emp_2: Self-explanatory/redundant
- Employment_Category: Whether the employee is Staff/Management/Labour
- Employment_Type: Whether the employee is Permanent/Contact_basic
- Religion: Employee's religion
- Birth year: Year of birth of the employee

Salary *(Size: 4.27 MB, Number of Attributes: 109, Number of Records: 9036)*
- Employee_No: The employee's identification number.
- Year: Year of salary payment
- Month: The month of salary payment
- Basic: Basic salary of permanent employees
- Basic per Day: Basic daily rate for contract basis employees
- Total_Working_Days: Number of total working days for the employee in a given month
- OT_Hours: Number of OT_hours(overtime hours) worked by the employee
- No pay days: Number of No pay leaves taken by the employee in a given month

Leaves *(Size: 53.6 kB, Number of Attributes: 6, Number of Records: 1018)*
- Employee_No: The employee's identification number.
- leave_date: The date on which the employee will be taking leave.
- Type: Whether the leave is Half/Ful day
- Applied Date: The day leave was recorded
- Remarks: Any additional remarks that the employee would like to add.
- apply_type: The type of application, such as "casual" or "annual".

Attendance *(Size: 16.4 MB, Number of Attributes: 10, Number of Records: 224057)*
- date: The date of the timesheet entry.
- project_code: The project code for the timesheet entry.
- out_date: The date on which the employee left the project.
- Employee_No: The employee's identification number.
- in_time: The time at which the employee started working.
- out_time: The time at which the employee stopped working.
- Hourly_Time: The number of hours worked.
- Shift_Start: The start time of the shift.
- Shift_End: The end time of the shift.

Holidays (*Size: 2 KB, Number of Attributes: 1, Number of Records: 121***)**
- Date: Date of the holiday

Data Preprocessing

Data quality issue: Missing values in birth_year in the employee data set
Action taken: Using average birth_year to fill in the missing values
Reasons: No highly co-related feature for imputing
Distribution is normally distributed

Data quality issue: Missing values in marital_status in the employee data set
Action taken: Using the classification model(Gradientboost) to predict the marital_status
Reasons: marital_status was highly correlated with birth_year

Data quality issue: The "Employee_ID" column in the salary dataset includes some IDs that are not present in the employee dataset
Action taken: Drop the relative rows
Reasons: Useless data.

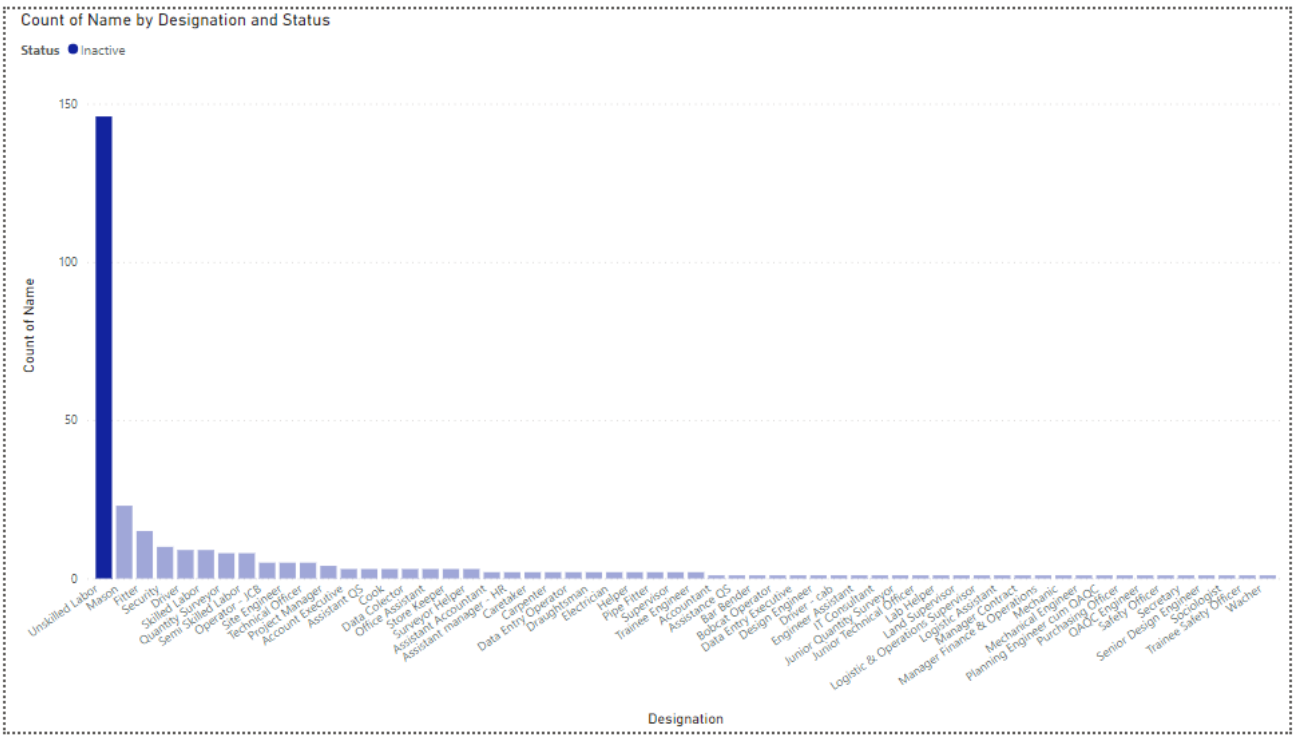
Data quality issue: Basic salary and Basic per day missing while Total Working Days are non-zero in the salary data set
Action taken: Replace Basic salary and Basic per day by the mean.
Reasons: The grouped by mean is the best estimate.

Data quality issue: Basic salary and Basic per day missing while Total Working Days are non-zero in the salary data set
Action taken: Replace Basic salary and Basic per day by the mean.
Reasons: The grouped by mean is the best estimate.

Data quality issue: Project_Code and Area missing in the salary set
Action taken: Replace Project_Code by the mode of employee and impute area
Reasons: Mode has the highest probability

Insights

Relationship between Employee Personal Data and Resignation



graph_02

According to graph_02 designations with the highest number of resignations in 2022

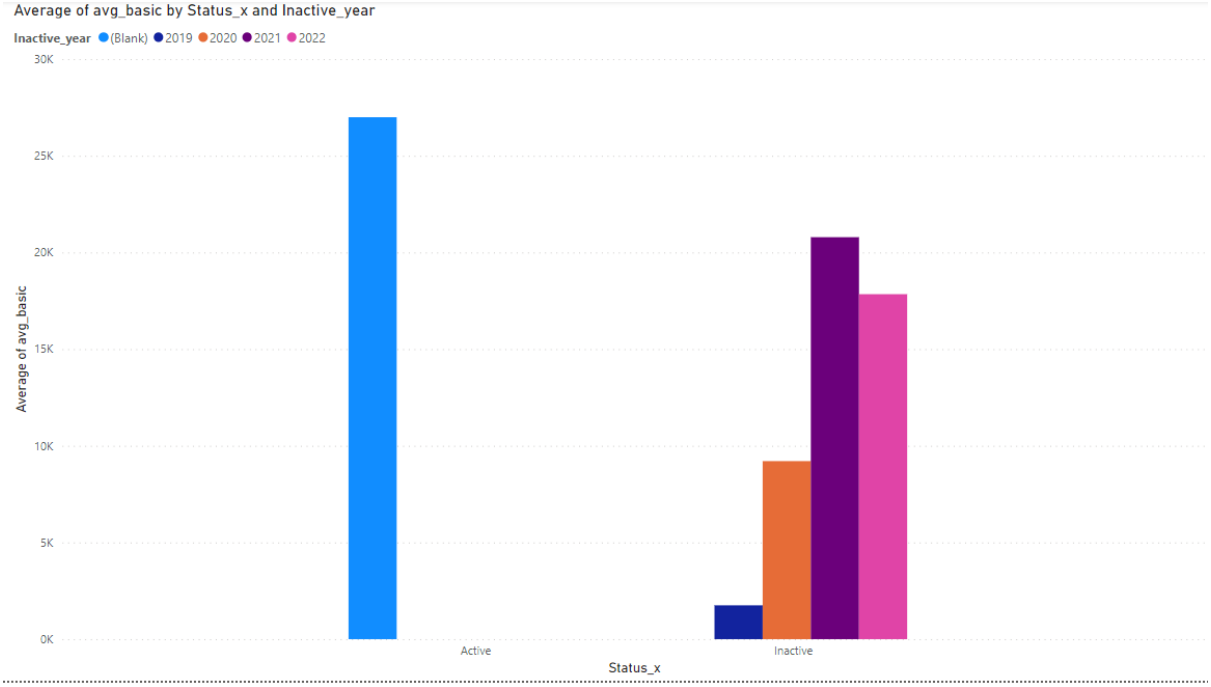
- Unskilled Labour
- Mason
- Fitter

Since Unskilled Labour, Mason, and Fitters are contract basis we could observe that contact basis employees have resigned the most in the year 2022.

Resignations of the year 2022 by contact basis|all employees according to Gender and Age.

Age	**Male**	**Female**
age<35	120|120	17|17
45>age>35	60|61	2|2
55>age>45	69|69	1|2
55<age	45|46	1|1

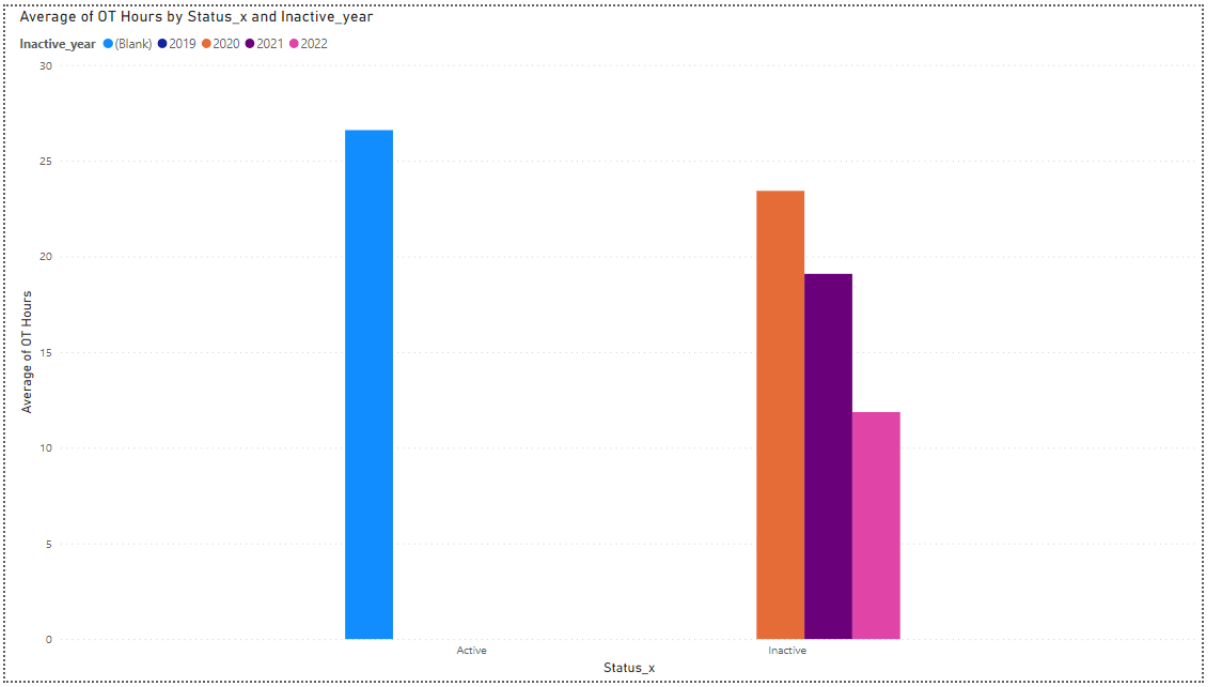
Relationship between Income and Resignation



graph_03

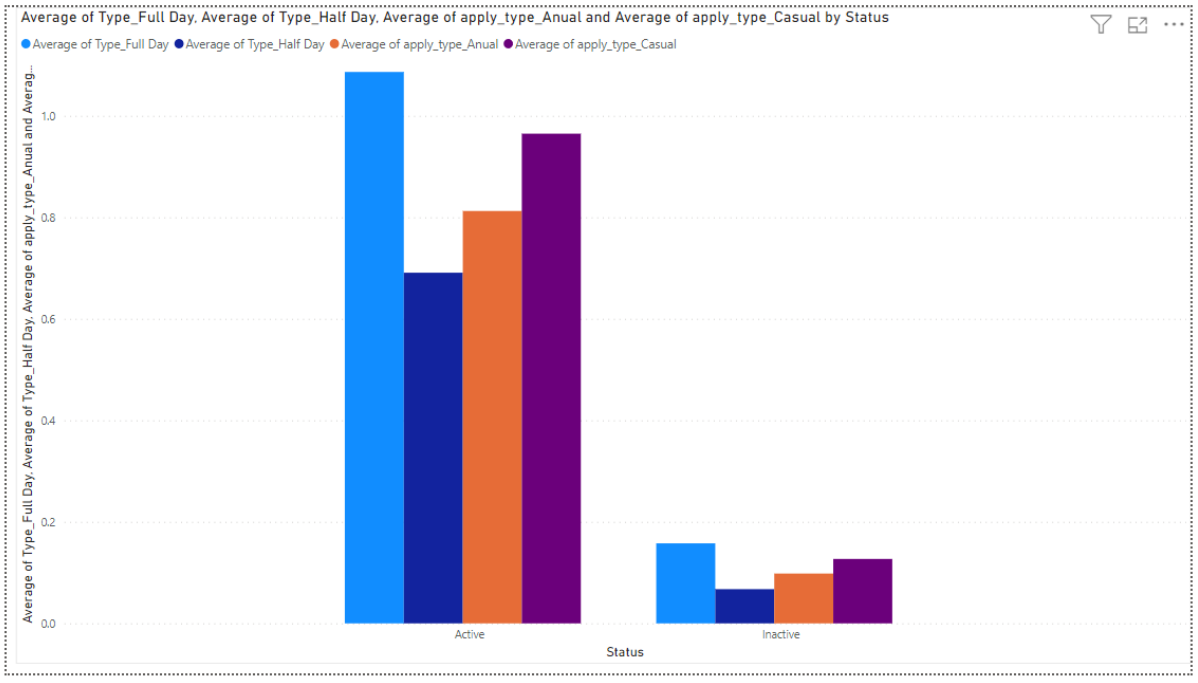
Graph_03 shows the average basic salary regarding the active/inactive status and resignation year. As it is observable in graph_03 average of ones who resigned in 2022 is less than the average basic of those who are still working.

Graph_04 shows the average OT hours regarding the active/inactive status and resignation year. Graph_04 gives the observation that the number of OT hours worked by the employees who resigned in 2022 is significantly less. The combination of less basic and fewer OT hours implies that the ones who resigned in 2022 have less monthly income compared to others. This drives us to the conclusion that **monthly income has a significant impact on employee resignation**. Less the monthly income it's more likely an employee to resign



graph_04

Relationship between leaves and resignation



graph_05

Graph_05 shows the number of leaves taken by active and inactive employees. This can be a clear indication of workload, stress, and living stress applied to employees.

However, the number of leaves taken by employees who resigned from the company in 2022 is relatively higher compared to the year 2021.

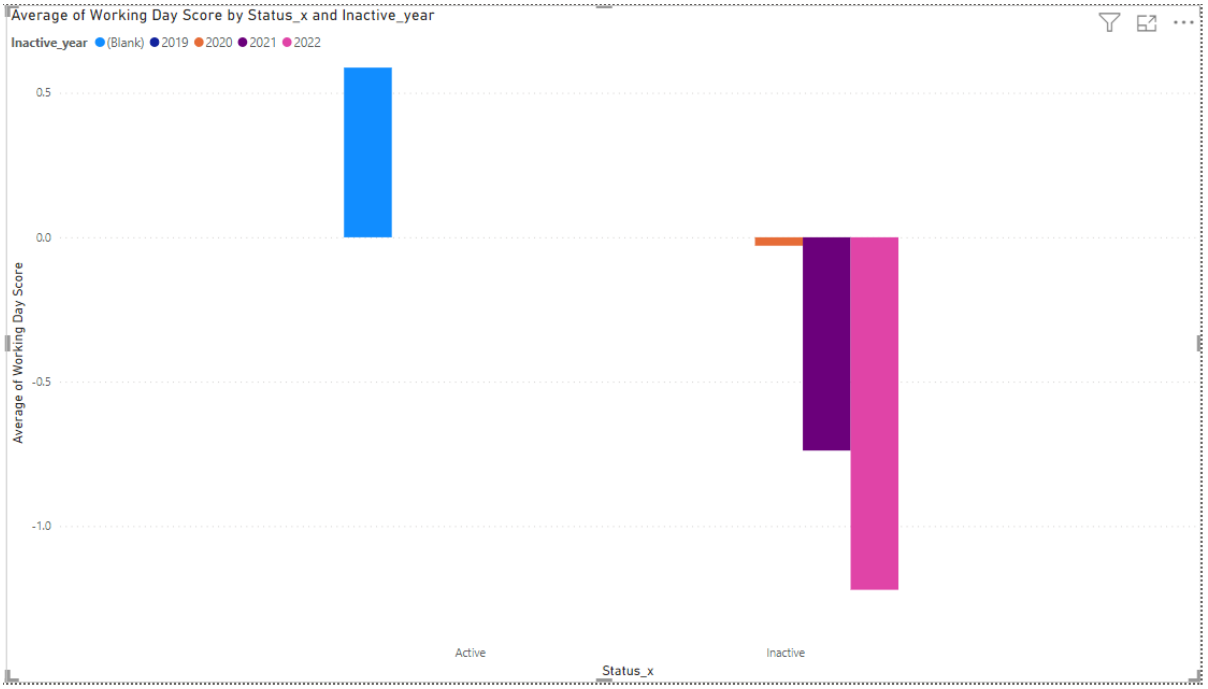
The number of resignations is 2022 and 2021

2021	204
2022	318

graph_06

Graph_05 shows the difference between leave type and year of resignation it's observable that ones who resigned in 2022 have taken a lot more leaves of all categories than ones who resigned in 2021.

Relationship between work Day score and resignation



graph_07

*__Work Day score__ measures employees' commitment to work recently giving higher weight to later and less weight to older. So according to graph_07 one who resigned in 2022 tends to have a negative workday score while active employees tend to have a positive workday score. This drives us to the conclusion of **loss of interest in work over time has an impact on resignation.**

Predictive Insight

Approach

- Use clustering(Unsupervised learning) to cluster employees into groups
- Identify high-risk clusters by calculating the ratio between inactive/all employees

Implementation

- Algorithm: AgglomerativeClustering (sklearn.cluster.AgglomerativeClustering)
- Number of clusters: 4
- Performance measure: silhouette_score - 0.63 (sklearn.matrics.sihouse_score)

Set of features used and description

Designation_ID	Identifier of employees designation
Year_of_Birth	Birth year of the employee
Year_joined	Year employee joined the company
Gender_Female	Is employee female
Gender_Male	Is employee male
Marital_Status_Married	Is employee Married
Marital_Status_Single	Is employee Single
Employment_Category_Labour	Is the employee a labor level
Employment_Category_Management	Is the employee a management-level
Employment_Category_Staff	Is the employee a staff-level
Employment_Type_Contarct Basis	Is the employee contract basis
Employment_Type_Permanant	Is the employee Permanent
isLate%	Percentage of employees being late to work
isEarly%	Percentage of employees being early
isHoliday%	Percentage of holidays employee has worked
avg_basic	Average basic of the employee
basic_latest_Salary	Latest basic of the employee
mean_working_days	Average worked days per month
latest_working_days	worked days in the last month
Working Day Score	Explained below

No Pay Days	Average monthly no-pay leaves
OT Hours	Average monthly OT hours

Working Day Score: A measurement used to calculate the trends in working days

formula:

$$\sum ((year * 13 - 20) + (month)/150) * Number\ of\ worked\ days\ in\ a\ month$$

Results of the clustering:

Cluster	**Status**	**No emplyees**
0	Active	120
0	Inactive	662
1	Active	19
1	Inactive	26
2	Active	94
2	Inactive	73
3	Inactive	3

Final conclusion

Cluster	Probability of resignation
0	**0.84**
1	**0.57**
2	**0.43**

Note: cluster 3 is omitted for this report but may have effect on new employees

If a employee belongs to cluster 0(/3) then the employee has higher probability of resigning.

Employee cluster allocation can be found here (employee_cluster_allocation.csv):
https://github.com/ImalshaD/DSProject200487B
(Includes codes for clustering (project.ipynb))