**Final Report**

*Marvelous Construction*

**PAHASARA G.G.V.**

**200440C**

Table of Contents

[1 Problem Overview 2](#_Toc140444820)

[2 Dataset Description 2](#_Toc140444821)

[3 Data Pre-processing 2](#_Toc140444822)

[3.1 Data Cleaning 2](#_Toc140444823)

[3.1.1 ‘employees’ dataset 2](#_Toc140444824)

[3.1.2 ‘salary’ dataset 4](#_Toc140444825)

[3.1.3 ‘attendance’ dataset 4](#_Toc140444826)

[3.1.4 ‘leaves’ dataset 4](#_Toc140444827)

[3.2 Data Reduction 5](#_Toc140444828)

[3.3 Data Integration 5](#_Toc140444829)

[3.4 Data Transformations 5](#_Toc140444830)

[4 Insights from Data Analysis 5](#_Toc140444831)

[4.1 Insight 1 5](#_Toc140444832)

[4.2 Insight 2 6](#_Toc140444833)

[4.3 Insight 3 6](#_Toc140444834)

[4.4 Insight 4 7](#_Toc140444835)

[4.5 Insight 5 8](#_Toc140444836)

# Problem Overview

Marvelous Construction, a major construction firm with 35 construction sites in different areas in Sri Lanka, is currently facing a significant challenge of high employee attrition. The task at hand is to understand the situation by analysis of data of Marvelous Construction as employee attrition is a serious situation. For this task a dataset containing employee details, attendance, leaves and salary is given.

In this report, the steps followed during dataset pre-processing and the five most significant insights gained from analyzing the data set is discussed.

# Dataset Description

Summary of information about the dataset is given below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| File | Shape | Columns | Number of Unique Employees | Notes |
| employees | (997, 19) | 'Unnamed: 0', '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' | 997 | Very Important dataset as this dataset contains our main objective of analysis, which is ‘Status’. The ‘Status’ is either ‘Active’ or ‘Inactive’ for an employee. |
| salary | (9035, 109) | 'Employee\_No', 'year', 'month', 'SiteNo', 'Area', 'Accomadation Allowance', 'Accommodation Allowance', 'Add. Allow. No Pay Deduction', 'Additional Allowance\_0', 'Additional Allowance\_2', ..., 'WSL No of Full Worked Days', 'WSL No of Half Days', 'WSL OT 1.5', 'WSL OT Hours 1.5', 'WSL OT Rate', 'WSL Total Earning', 'WSL Total Working Days', 'WSLOther Allowance Rate', 'Working Days - Full', 'Working Days - Half Day' | 1598 | The intersection with employees on ‘Employee\_No’ results in a dataset containing 719 unique employees. |
| attendance | (224057, 10) | 'id', 'project\_code', 'date', 'out\_date', 'Employee\_No', 'in\_time', 'out\_time', 'Hourly\_Time', 'Shift\_Start', 'Shift\_End' | 1883 | The intersection with employees on ‘Employee\_No’ results in a dataset containing 742 unique employees. |
| leaves | (1018, 6) | 'Employee\_No', 'leave\_date', 'Type', 'Applied Date', 'Remarks', 'apply\_type' | 70 | The intersection with employees on ‘Employee\_No’ results in a dataset containing 57 unique employees. |

# Data Pre-processing

## Data Cleaning

### ‘employees’ dataset

* There were no duplicate entries in the ‘employees’ dataset.
* The ‘Employee\_No’ is a unique field meaning there are no multiple entries with the same ‘Employee\_No’
* There were 67 missing values in the ‘Marital\_Status’ field.
* There were 103 entries with the invalid value ‘0000’ for ‘Year\_of\_Birth’ field.
* There were 183 and 105 entries with the invalid value ‘0000-00-00’ for ‘Date\_Resigned’ and ‘Inactive\_Date’ fields, respectively.
* There were data inconsistencies in ‘Title’, ‘Gender’ and ‘Marital\_Status’ Fields.

#### Resolving data inconsistencies in ‘Title’, ‘Gender, and ‘Marital\_Status’ fields

* There were 10 entries with ‘Gender’ as ‘Male’ and ‘Title’ as ‘Ms’ or ‘Miss’.
* There were 4 entries with ‘Gender’ as ‘Female’ and ‘Title’ as ‘Mr’.
* There was an entry with ‘Title’ as ‘Miss’, ‘Gender’ as ‘Female’ and ‘Marital\_Status’ as ‘Married’.

To resolve the above issues, the ‘Gender’ field is assumed to be correct. The ‘Name’ field cannot be utilized for resolving the above issues as most of the names does not make it clear to which ‘Gender’, ‘Title’ and ‘Marital\_Status’ a person belong to. Therefore, the other field are transformed accordingly as follows on the assumption that ‘Gender’ is correct.

* The entries with ‘Gender’ as ‘Male’ and ‘Title’ as ‘Ms’ or ‘Miss’, the title is changed to ‘Mr’.
* The entries with ‘Gender’ as ‘Female and ‘Title’ as ‘Mr’, the title is changed to ‘Ms’ and kept the ‘Marital\_Status’ as it is.
* In the entry with ‘Title’ as ‘Miss’, ‘Gender’ as ‘Female’ and ‘Marital\_Status’ as ‘Married, the ‘Title’ is changed to ‘Ms’.

#### Resolving the invalid values in ‘Date\_Resigned’ and ‘Inactive\_Date’ fields

In almost all entries, the ‘Date\_Resigned’ is the same as ‘Inactive\_Date’. If the ‘Status’ is ‘Active’, meaning the employee is still in service, then that employee does not have a ‘Date\_Resigned’ or a ‘Inactive\_Date’, shown as a ‘\N’ in dataset. By using the above facts, the issues are solved as follows:

* ‘Inactive\_Date’ is copied to ‘Date\_Resigned’ where ‘Date\_Resigned’ is ‘0000-00-00’ or ‘\N’ and ‘Status’ is ‘Inactive’
* Replaced with ‘\N’ where ‘Status’ is ‘Active’ for both ‘Date\_Resigned’ and ‘Inactive\_Date’.

#### Imputing the missing values of ‘Marital\_Status’ field

* ‘Marital\_Status’ is a binary value that is either ‘Married’ or ‘Single’. To impute the binary valued ‘Marital\_Status’, a classification algorithm, DecisionTreeClassifier is used.
* 'Employee\_Code', 'Name', 'Religion\_ID', 'Designation\_ID', 'Date\_Resigned', 'Inactive\_Date', 'Reporting\_emp\_1', 'Reporting\_emp\_2' fields are droped from the dataset for the classification. The ‘Employee\_Code’ is yet another unique value for an employee as the ‘Employee\_No’. ‘Name’ cannot be used for classification. ‘Religion\_ID’ and ‘Designation\_ID’ fields are just label encoding fields for ‘Religion’ and ‘Designation’. ‘Date\_Resigned’ and ‘Inactive\_Date’ contains ‘\N’ values which cannot be converted to a meaningful numerical representation. And lastly the 'Reporting\_emp\_1' and 'Reporting\_emp\_2' fields are mostly ‘\N’ which are not meaningful.
* ‘Date\_Joined’ field is converted to number of days since the epoch to the ‘Date\_Joined’, and inserted to a new column called ‘Date\_Joined\_Days’.
* Applied one-hot encoding to the fields, 'Title', 'Gender', 'Status', 'Employment\_Category', 'Employment\_Type', 'Religion' and 'Designation'.
* Droped the fields used to generate new fields for classification. ('Title', 'Gender', 'Date\_Joined', 'Status', 'Employment\_Category', 'Employment\_Type', 'Religion', 'Designation')
* For the invalid value ‘0000’ in ‘Year\_of\_Birth’, those values are replaced by the median of the ‘Year\_of\_Birth’ field as the ‘Year\_of\_Birth’ field has a skewed distribution.
* Then, hyperparameters for the DecisionTreeClassification is chosen via a random search of 150000 fits with 5-fold cross validation.
* The DecisionTreeClassification gives about 86.4% classification accuracy for ‘Marital\_Status’.

#### Imputing the invalid values of ‘Year\_of\_Birth’ field

* The ‘Year\_of\_Birth’ is a discrete value. There are not enough data to take ‘Year\_of\_Birth’ as a categorical values and use classification for imputation. Therefore, a regression model is used. Since the exact trend is difficult to recognize with the mostly categorical fields in ‘employees’ dataset, a DecisionTreeRegression model is used to impute the ‘Year\_of\_Birth’.
* The same dataset prepared to impute ‘Marital\_Status’ is used for this step with the imputed ‘Marital\_Status’ values entered.
* Then, hyperparameters for the DecisionTreeRegressor is chosen via a random search of 150000 fits with 5-fold cross validation.
* The DecisionTreeClassifier is able to give a root mean squared values of about 10.18 for ‘Year\_of\_Birth’ field.
* There was only one outlier almost touching the lower fence of box plot of ‘Year\_of\_Birth’ field. Therefore, that outlier was ignored.

### ‘salary’ dataset

* There are no duplicate entries in ‘salary’ dataset
* There are 45 missing values in ‘SiteNo’ field.
* There are 45 entries with invalid value ‘\N’ for ‘Area’.
* There are no other data inconsistencies other than the above-mentioned issues.

#### Imputing missing values and invalid entries of ‘SiteNo’ and ‘Area’ fields

The following logic is used for imputing missing values and to resolve invalid entries. It is important to note that all the entries with ‘SiteNo’ missing have the invalid ‘Area’ value, which is ‘\N’.

1. If the ‘SiteNo’ is missing in an entry, get all the entries with the same ‘Employee\_No’ and ‘SiteNo’ not missing (or ‘Area’ not invalid).
2. If the entry list returned by step 1 is empty, go to step 5.
3. Look for an entry that is closest in time to the entry with missing value.
4. There must be an entry since the entry list returned by step 1 is not empty. Replace the ‘SiteNo’ and ‘Area’ of entry with missing value with those with the selected entry. Then go to step 7.
5. Look for entries that are closest in time with the entry with missing value.
6. Replace the ‘SiteNo’ and ‘Area’ of entry with missing value with those with the modes of entries returned by step 5.
7. End of procedure.

### ‘attendance’ dataset

* There were no duplicate entries in ‘attendance’ dataset
* There are 15 inconsistence values for ‘out\_time’. Representing mid-night as 0:00:00 will be more consistence than 24:00:00. Moreover, there were ‘out\_time’ entries as 24:50:00 which is inconsistence.
* There were 10 inconsistence values for ‘out\_date’ field. In the event that the ‘out\_time’ is 0:00:00(or 24:00:00), some of the ‘out\_date’ are of the same day as ‘date’ while others are the next day to ‘date’.
* There were 15 entries with invalid value ‘\N’ for ‘Hourly\_Time’ field.
* In the row with index 142203, there is an invalid entry as ‘0000-00-00’ for ‘out\_date’.
* There are no other data inconsistencies other than the above-mentioned issues.

#### Resolving data inconsistency issues mentioned above

* ‘out\_time’ starting from 24 hours and after are converted to standard form (e.g.: - 24:50:00 -> 00:50:00)
* For inconsistence ‘out\_date’ values mentioned above, replace ‘out\_date’ with the next day from ‘date’.
* There are no inconsistence and missing ‘in\_time’ and ‘out\_time’ at this point. Simply subtract ‘out\_time’ from ‘in\_time’ and replace them with invalid ‘\N’ values in ‘Hourly\_Time’.
* Manually change the ‘out\_date’ of index 142203 to ‘3/1/2022’ as the ‘date’ is ‘2/28/2022’ and the ‘out\_time’ is ‘0:45:03’.

### ‘leaves’ dataset

* There were some duplicate entries in the ‘leaves’ dataset.
* There were 245 missing values.
* There are no other issues other than the above-mentioned issues.

#### Resolving issues mentioned above

* Droped the duplicate entries from ‘leaves’ dataset.
* Replaced null values with ‘\N’ placeholder.

## Data Reduction

* Remove unwanted or redundant fields from ‘employees’ dataset: 'Employee\_Code', 'Religion\_ID', 'Designation\_ID', 'Reporting\_emp\_1', 'Reporting\_emp\_2'

## Data Integration

* ‘employees’, ‘salary’, ‘attendance’ and ‘leaves’ datasets are combined to a one model in Power BI with one-to-many relationships between ‘employees’ dataset and the other three datasets by the ‘Employee\_No’ field.

## Data Transformations

* When importing the datasets to Power BI, the field with date and time value type are separated to year, month, day, hour, minute for easy use for visualization.
* Refer the following links for more information on data cleaning.
* ‘employees’ dataset cleaning – [Jupyter Notebook](https://drive.google.com/file/d/1NjPotpdeu-tyuhrP1HvSBKe1CFYzAOC_/view?usp=sharing) | [python script](https://drive.google.com/file/d/1285tL99IOfQTEoWlgt4xIRi8DMVwTX75/view?usp=sharing)
* ‘employees’ dataset cleaning – Jupyter Notebook | python script
* ‘salary’, ‘attendance’ and ‘leaves’ datasets cleaning – [Jupyter Notebook](https://drive.google.com/file/d/1MJxlB2s8C5mTTQfVMz8faH5gLU02PFOY/view?usp=sharing)
* ‘salary’, ‘attendance’ and ‘leaves’ datasets cleaning – Jupyter Notebook
* Integrated, transformed datasets – [Power BI file](https://drive.google.com/file/d/1CWKjygsw1Y11DdPXGYuUjDD204S7nOe3/view?usp=sharing)
* Integrated, transformed datasets – Power BI file

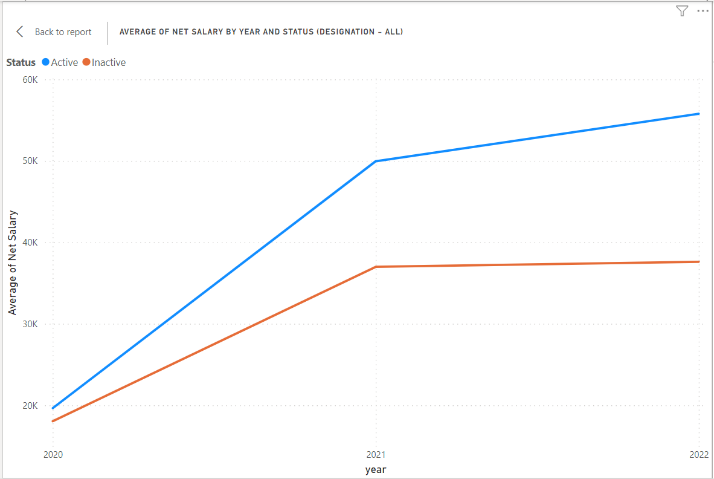
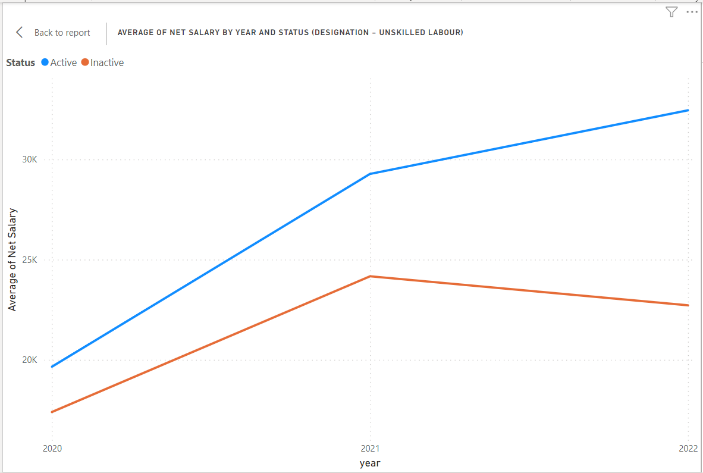
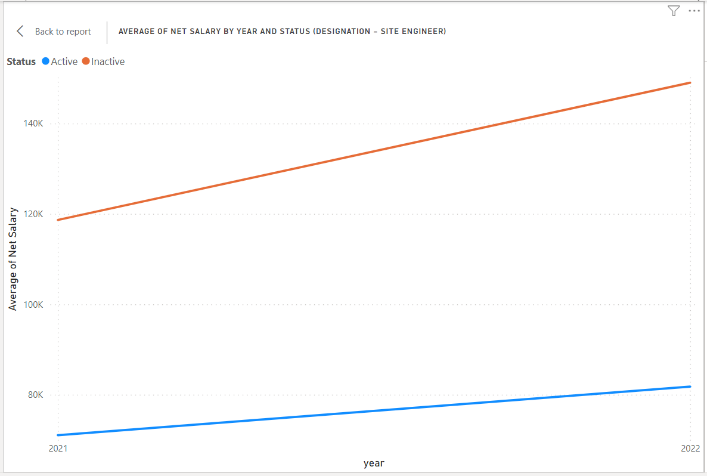
# Insights from Data Analysis

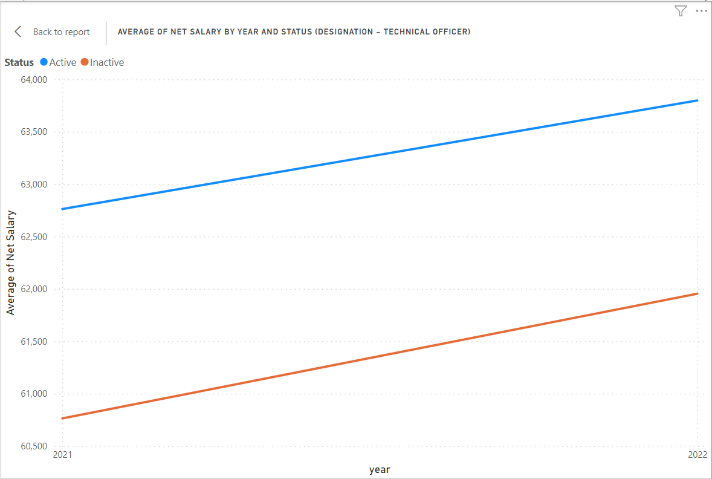
## Insight 1

**Overall, the employees who get a higher net salary tend to remain employed while employees who get a low net salary tend to resign. This trend is true for most of the designations (Unskilled Labor,Labour, Technical officer, Supervisor, Store Keeper, …) while a few designations (Site Engineer, Project Manager) deviate from this common trend.**

**Approach:**

Average net salary is graphed against year for employee status, active and inactive choosing all employees or employees of a specific designation.





The last graph shows a case of deviation where the designation is Site Engineer. These deviations may be the result of unique circumstances of those designations.

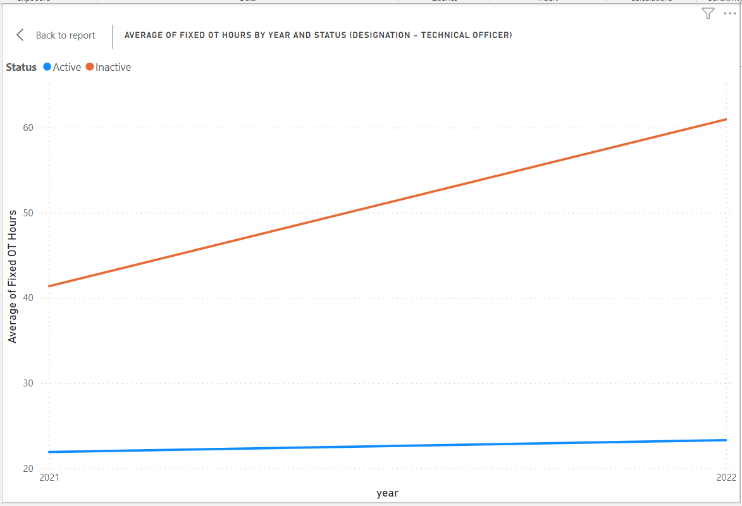
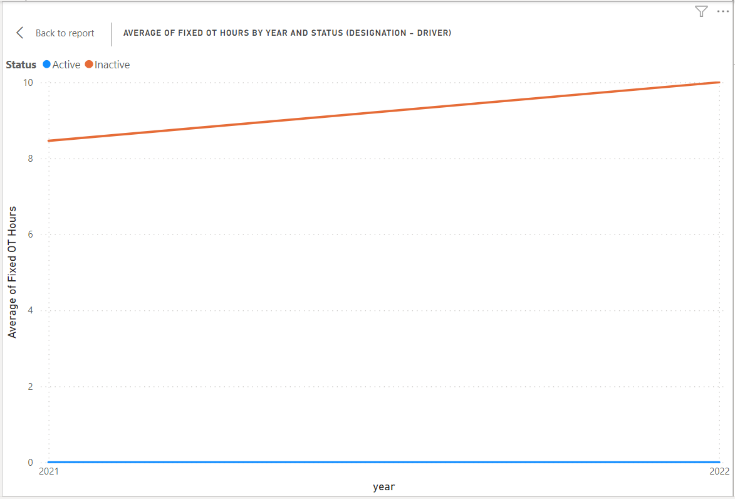
The last graph show a case of deviation where the designation is Site Engineer. This deviations may be the result of unique ciecumetances of those designations.

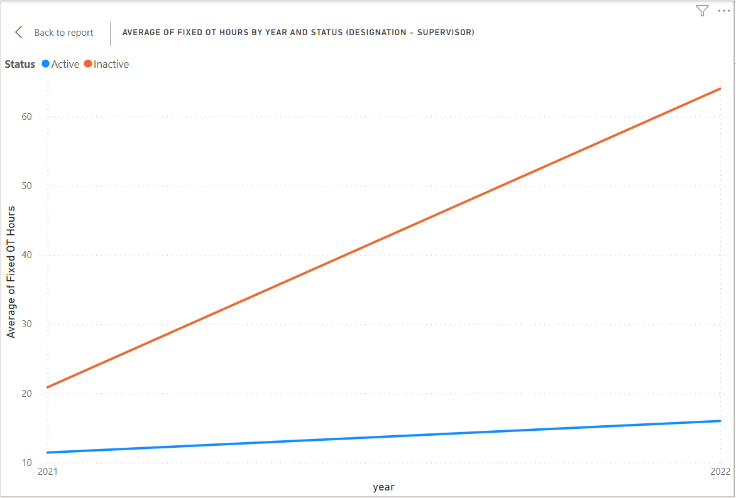
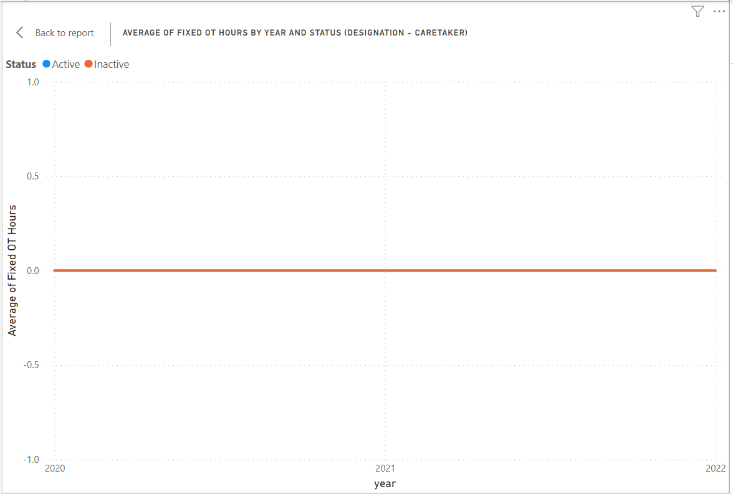
## Insight 2

**Employees who have higher fixed OT hours tend to resign while employees who have lower fixed OT hours tend be remain employed. This trend is true for almost all the designations except few including the designations without any fixed OT hours assignment for employees.**

**Approach:**

Average Fixed OT Hours is graphed against year for employee status, active and inactive choosing employees of a specific designation.





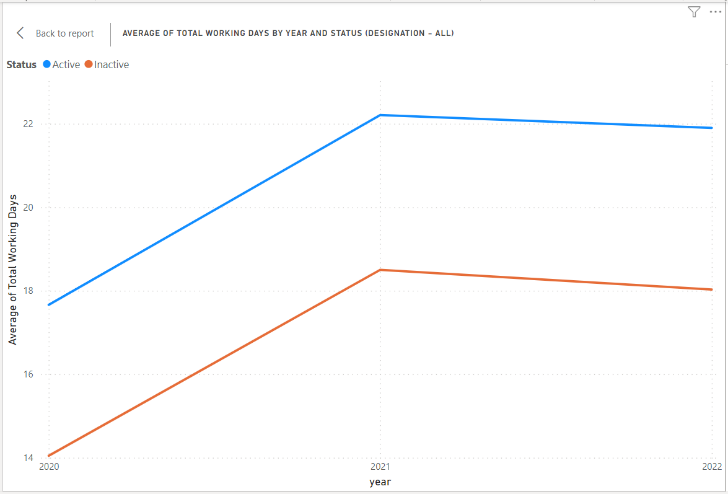
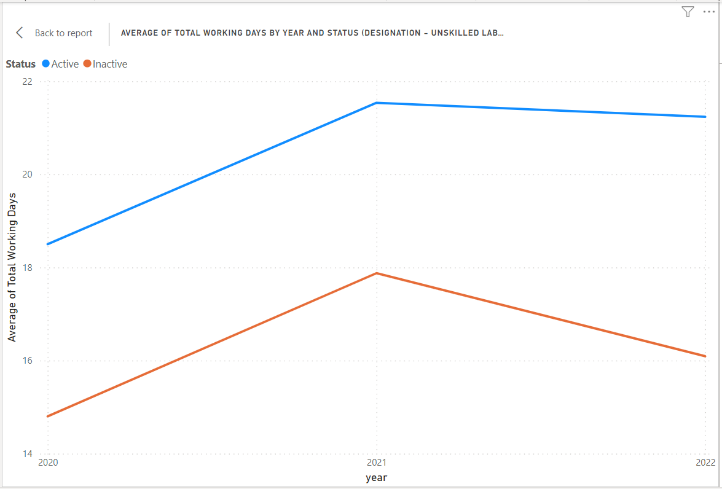
The last graph shows a case of deviation where the designation is Caretaker. In that designation, no Fixed OT Hours for employees.

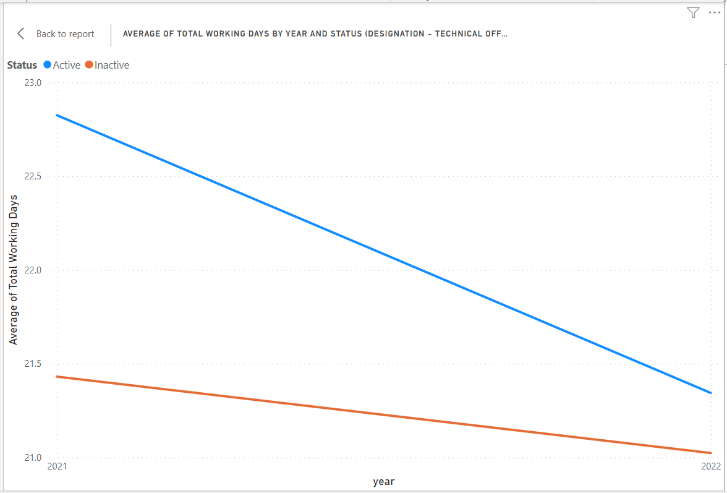
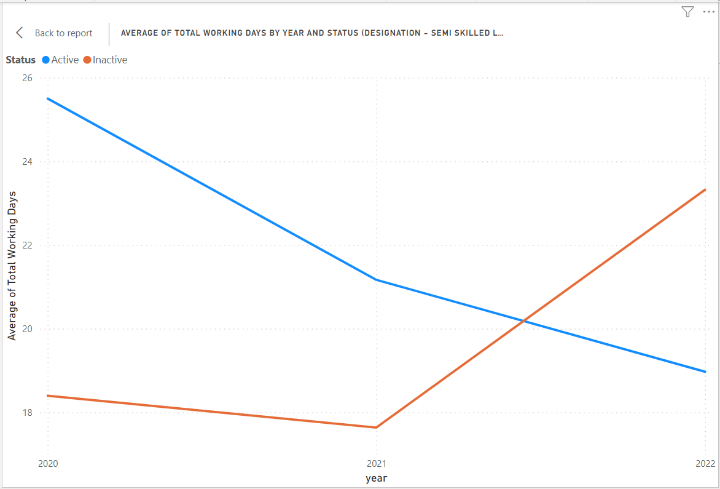
Insight 3

**Overall, the employees with higher attendance tend to remain employed while employees with lower attendance tend to resign. This trend is true for most of the designations (Unskilled Labor, Technical officer, Supervisor, Quantity Surveyor, …) while a few designations (Semi Skilled Labor, Store Keeper) deviate from this common trend. Therefore, employee attendance can be used as a factor to determine the probability of an employee is resigning in near future.**

**Approach:**

Average of Total Working Days is graphed against year for employee status, active and inactive choosing all employees or employees of a specific designation.





The last graph shows a case of deviation where the designation is Semi Skilled Labor. These deviations may be the result of unique circumstances of those designations.

## Insight 4

**The employees with a net salary of about 70K or more do not get an Attendance Allowance\_0. Single employees who get an Attendance Allowance\_0 of about 3.5K or more tend to remain employed while single employees who get an Attendance Allowance\_0 of less than 3.5K tend to resign. For married employees, the decision of resigning or remaining employed does not depend on Attendance Allowance\_0.**

**Approach:**

A scatter plot of Attendance Allowance\_0 vs Net Salary is created (Figure 1). Observe how no employees with net salary of more than 70K does not have an Attendance Allowance\_0. Filter data to have a Net Salary and Attendance Allownce\_0 of more than 0. Then filter the scatter plot for married employees (Figure 2) and single employees (Figure 3). Observe how there is no distinct clusters within Figure 2. Therefore, no noticeable impact on the employee status by having an Attendance Allowance\_0 for married employees. Observe how there are two distinct clusters within Figure 3. A line with equation, Attendance Allowance\_0 = 3.5K will divide the clusters for a very high accuracy. Therefore, single employees who get an Attendance Allowance\_0 of more than 3.5K tend to remain employed with high probability.

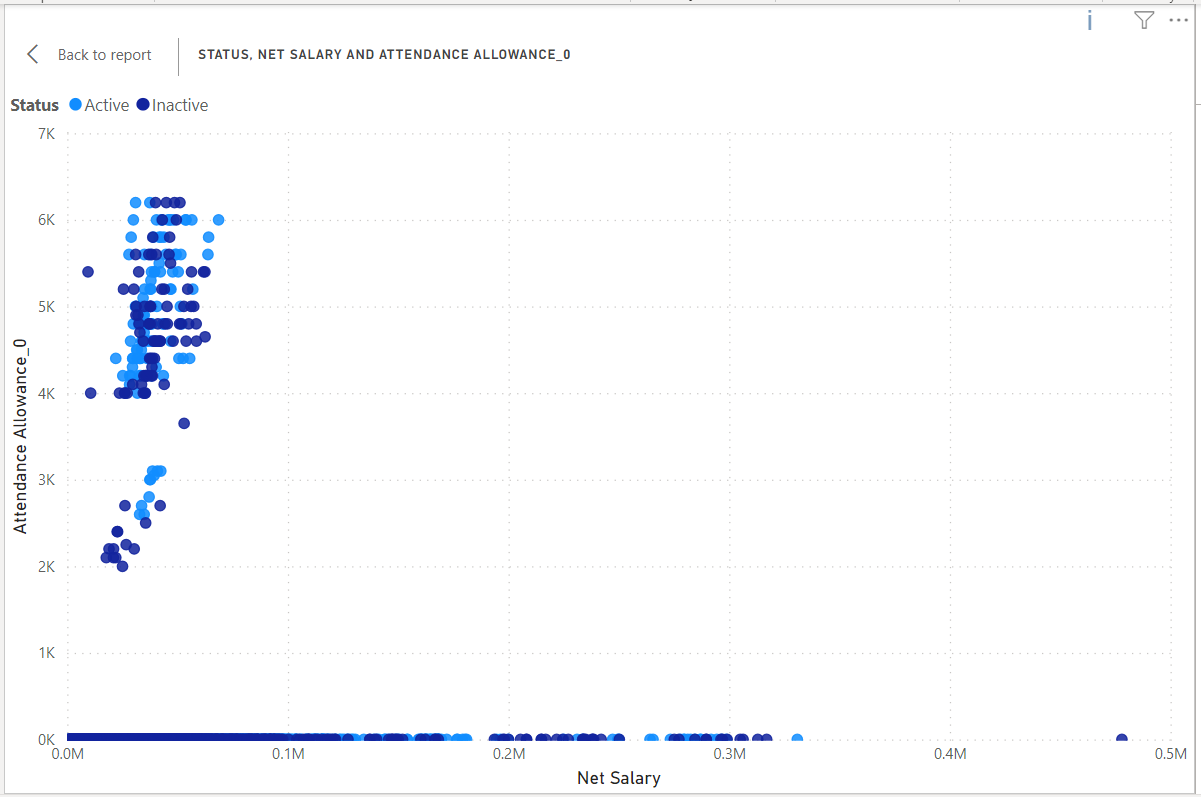


Figure 1

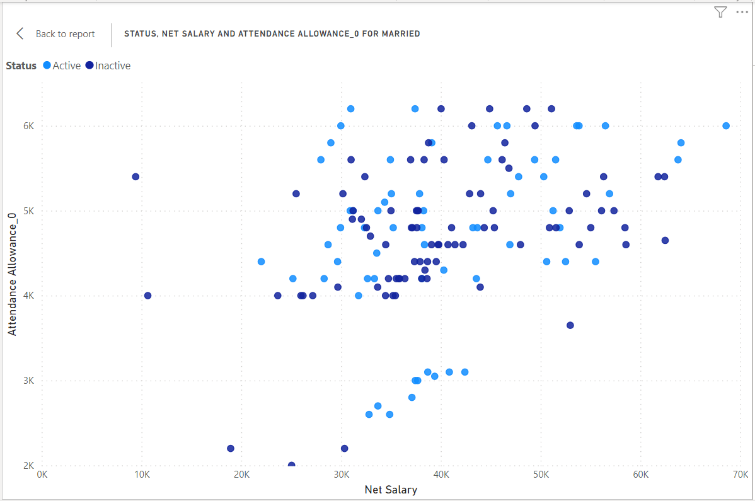
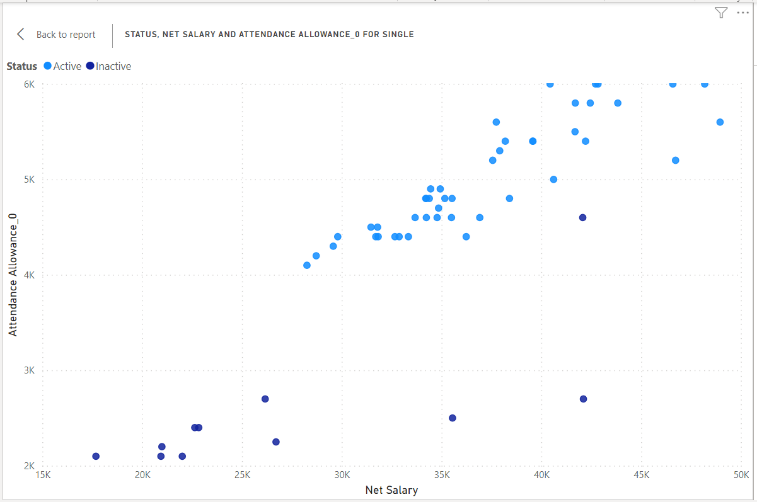


Figure 3

Figure 2

## Insight 5

**Employees who get a higher Attendance Allowance\_2 tend to remain employed than the employees who get a higher Accommodation Allowance. Employees tend to resign if they do not get a high Attendance Allowance\_2 irrespective of the Accommodation Allowance they get (high or low). Therefore, to encourage employees to remain employed, an Attendance Allowance\_2 should be given to employees.**

**Approach:**

A scatter plot of Average of Attendance Allowance\_2 by months vs Average of Accommodation Allowance by months is created (Figure 1). Observe that there are two distinct clusters. One cluster getting a considerable Attendance Allowance\_2 while the other cluster is getting a considerable Accommodation Allowance\_2. The employees within the cluster that gets a considerable Attendance Allowance\_2 remain employed while the employees who get a considerable Accommodation Allowance have resigned. Therefore, to encourage employees to remain employed, an Attendance Allowance\_2 of a considerable amount is recommended.

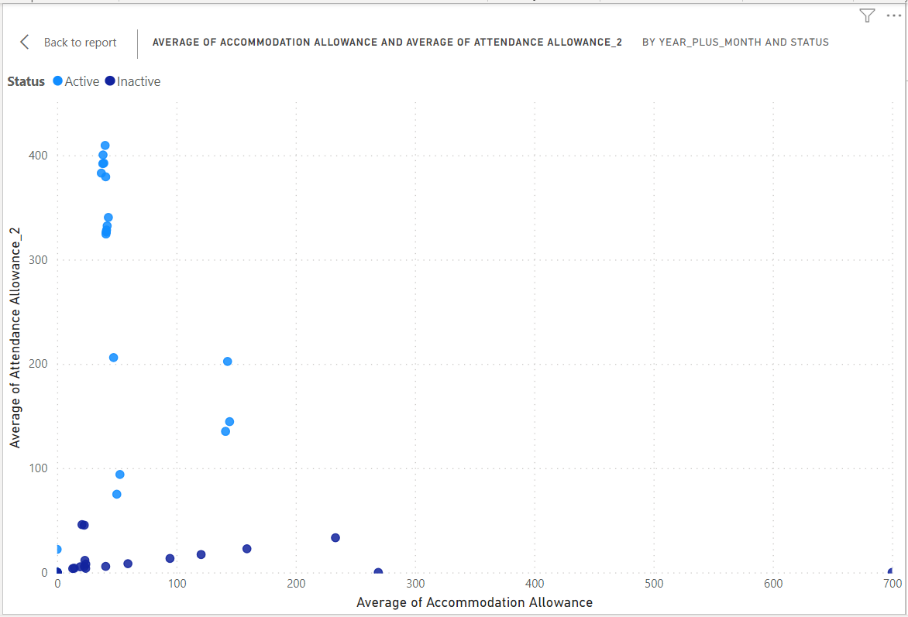


Figure 1