Problem Overview

- **Marvelous** Construction is a major construction firm with 35 construction sites in different areas in Sri Lanka. The Human Resources department of Marvelous Construction has recently noticed that **many employees are resigning**. For this problem, by analyzing the given data, the data scientist is supposed to give insights about this matter. In this case, the data scientist is supposed to search for the connections between given data to get insight into the causes of leaving employees. The main problems to give solutions are, what are the main reasons for employees to resign from the company, and what is supposed to be done to stop the resigning of the employees.

Dataset Description

- The given data set includes the main six files including data of employees and related data descriptions.

File Name	Number of Records	Remarks
employee	631	This consists of the main private details of the Employees. After pre-processing it is named as employee_preprocess_200638P file, and it is used to get insights.
attendance	60354	This consists of the records of the attendance of the employees.
salary	2632	This consists of the salary details of the employees.
leaves	237	This consists of the records of the leaves of employees.
salary_dictionary	***	This consists of the main components of the salary of employees.
holidays	***	This consists of the holidays in the company.

- In search for insights into the given problem mainly used datasets are **employee, attendance, salary, and leaves.**

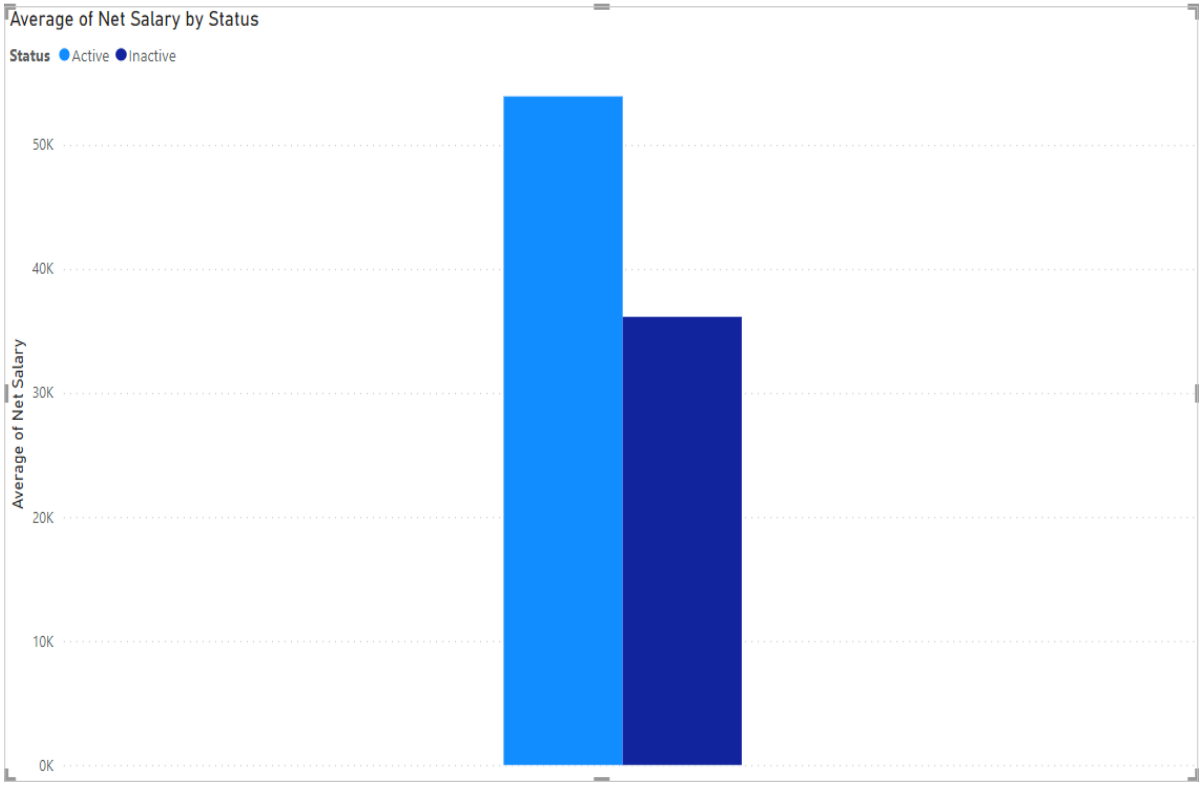
Data Preprocessing

- In the data preprocessing step, in the employee.csv file, **'Year_of_Birth'** and **'Marital_Status'** were imputed using the **decision tree regression and classification**.
- First, null/blank values of '**Title**' were filled out by checking the values of **'Marital_Status'** and **'Gender'**.
- Before using the regressor or classifier, **the 'Employment_Type', 'Employment_Catagory', 'Gender', 'Title', 'Status', 'Designation', 'Religion',** and **'Date_Joined'** was encoded using a label encoder.

- After that, using the **'Employment_Type', 'Employment_Catagory', 'Gender', 'Title', 'Status', 'Designation', 'Religion',** and **'Date_Joined'** decision tree classifications used to predict the **'Marital_Status'** and decision tree regression is used to predict **'Year_of_Birth'**.
- When using the decision tree, **Randomized Search** is used to get the highest accuracy from the model, the classifier, and the regressor.
- **All null valued records in salary.csv and attendance.csv files were dropped out of consideration.**
- When analyzing the data sets, a new column was created to add both '**Marital_Status**' and '**Gender**'. For example, if someone is married and male it will show that as '**Married_Male**'.
- When considering analysis, the '**leaves.csv'** file's data was inner joined with the '**employees_preprocessed_200638P.csv**' file and with the '**attendance.csv**' file.

Insights of Data Analysis

- ### Average Net Salary by Status



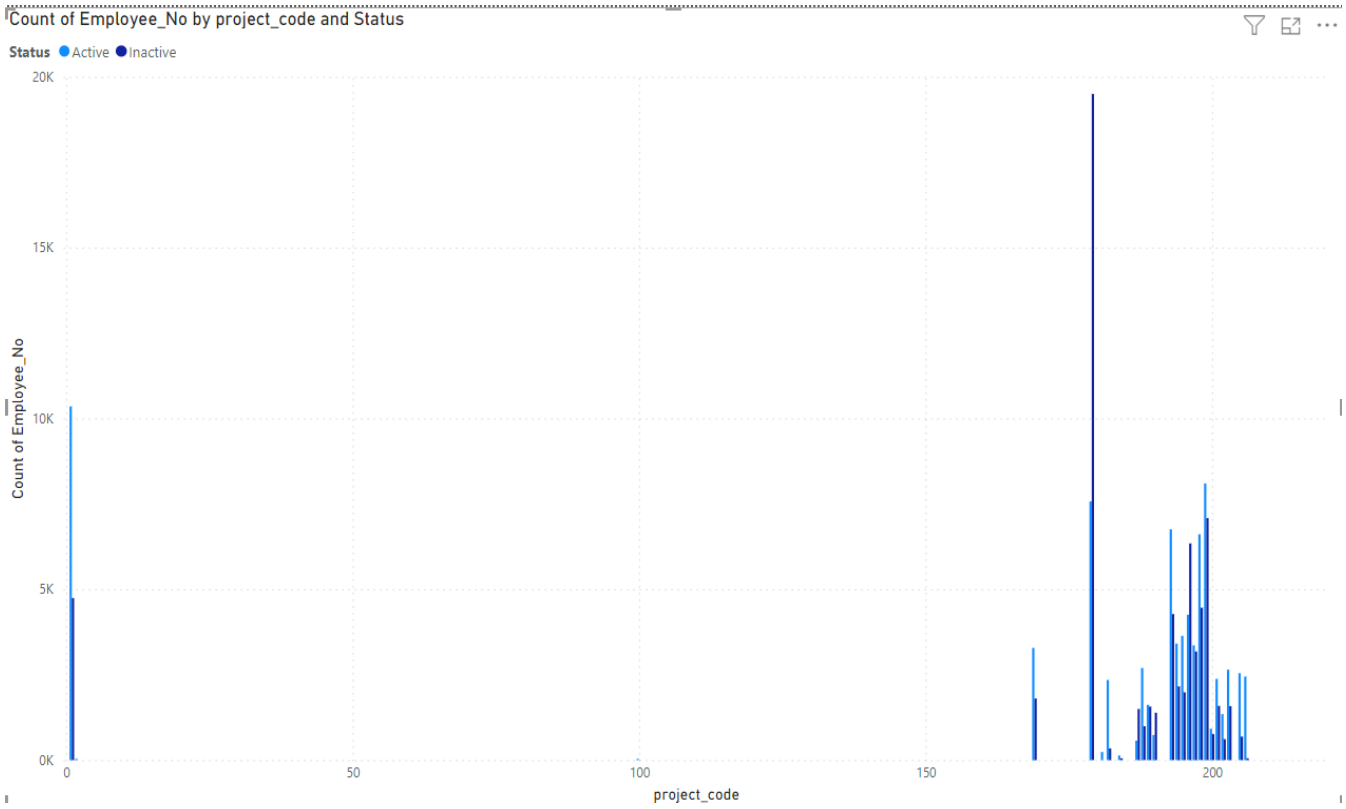
This bar graph shows how much average net salary was distributed to active and inactive employees. When considering active employees, they get a Rs.53915.63 average net salary; when considering inactive employees, they get a Rs.36138.50 average net salary. That means most of the employees who got resigned got an average net salary of less than the active employees in the company. There is a

difference of 17777.13 rupees. Through this analysis, we can get the idea that **if the employees get less amount of net salary tends to leave the company.**

By considering how they got paid, we can analyze by looking at the data the inactive employees tend to get lower salaries and they got their basic salaries less than others. This can be a major reason for them to leave the company.

Increasing the basic salaries that will cause to increase in the net salaries which will cause to reduce the leaving number of company employees.

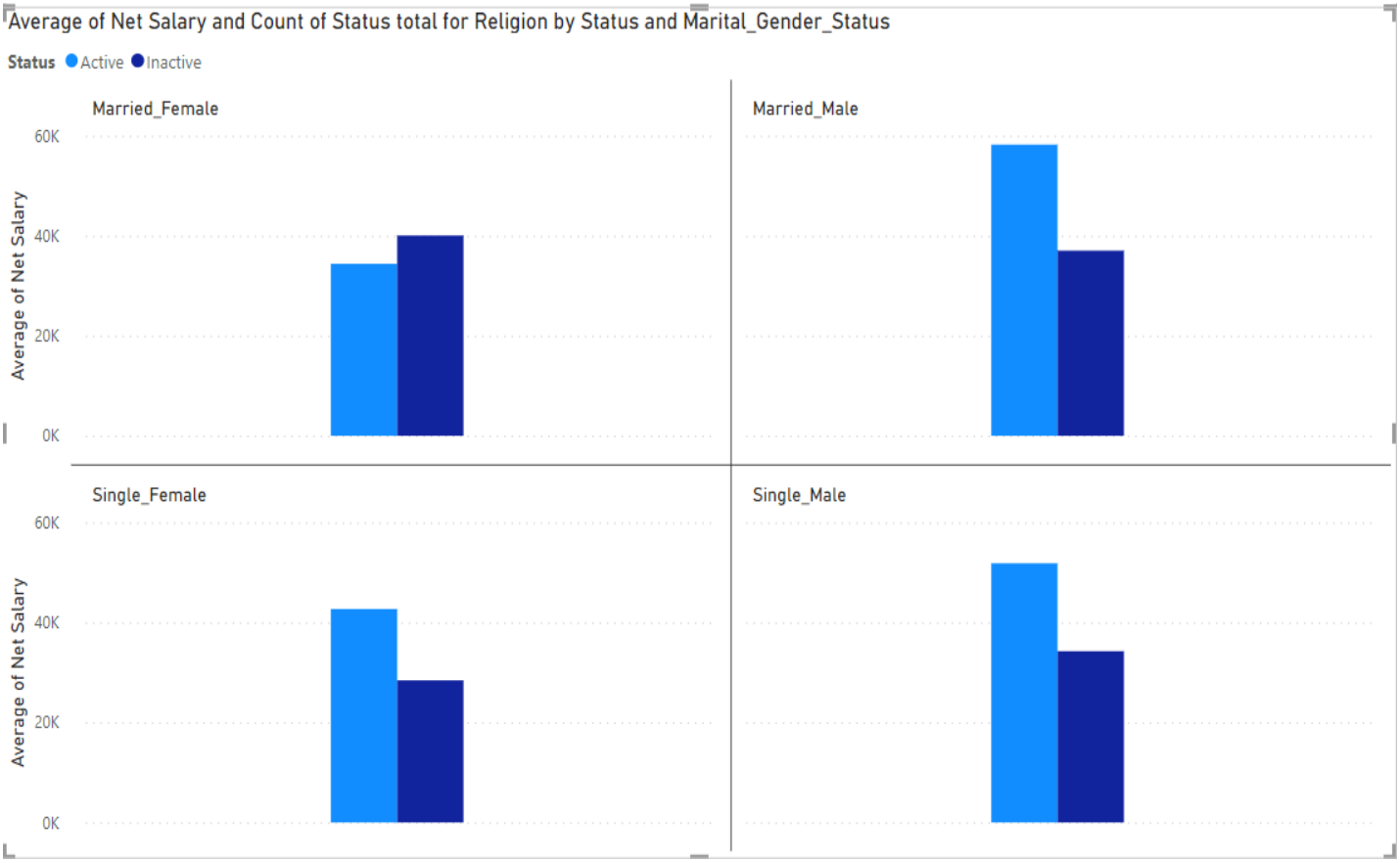
- ## Counts Of Employees by Status & Project Code



By considering the given bar graph we can get **the project codes that the greatest number of employees resigned**. In the graph, the x-axis is representing the Project code, and the y-axis represents the count of employees in that project both inactive and active. Active employees are represented by light blue bars and inactive employees by dark blue bars.

When considering this graph there are 4 projects there are more inactive people than active people, their codes are **179, 187, 190, and 196**. In those projects, there are 44%, 45%, 32%, and 20% percentages are greater the inactive employees than active employees respectively. There are considerable inactive employee percentages in projects that have project codes, **189, 200, and 201**. In these projects, employee resignation is a major issue and must be considered.

- **AVERAGE NET SALARY BY STATUS, MATERIAL_GENDER_STATUS, & STATUS**

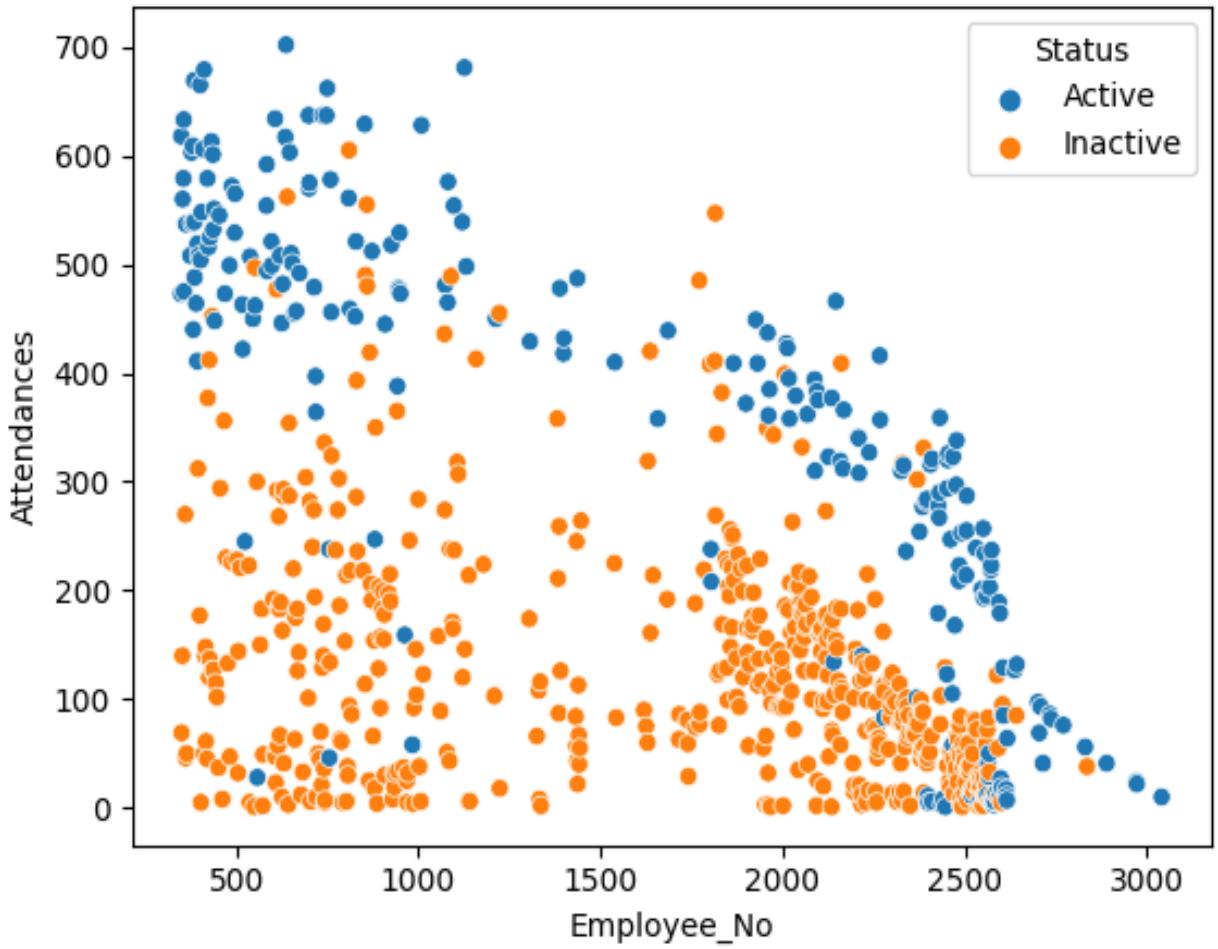


In the given bar graph, the average net salary, status, marital status, gender, and status of the employees were considered. The y-axis represents the average net salary of employees while the light blue bars show active employees the dark blue bars show inactive employees. The 4 categories show the married females, married males, single females, and single males respectively by each graph in each square.

When considering the graphs here, there is the same pattern **other than in married females**. Other than in the married female graph, all the inactive employees got low net salaries than active employees which may cause them to leave the company. But considering the married females, active employees' salaries are lower than inactive employees. This may cause because of their social life, most of the time females tend to leave their job after they got married without considering the salary. Therefore, it is **good to consider these marital and gender categories in employees**.

When considering overall data, **inactive employees got lesser amounts of salaries than active employees**. But after looking at the marital and gender status with regards to activeness, that must be analyzed.

- **ATTENDANCES VS STATUS**

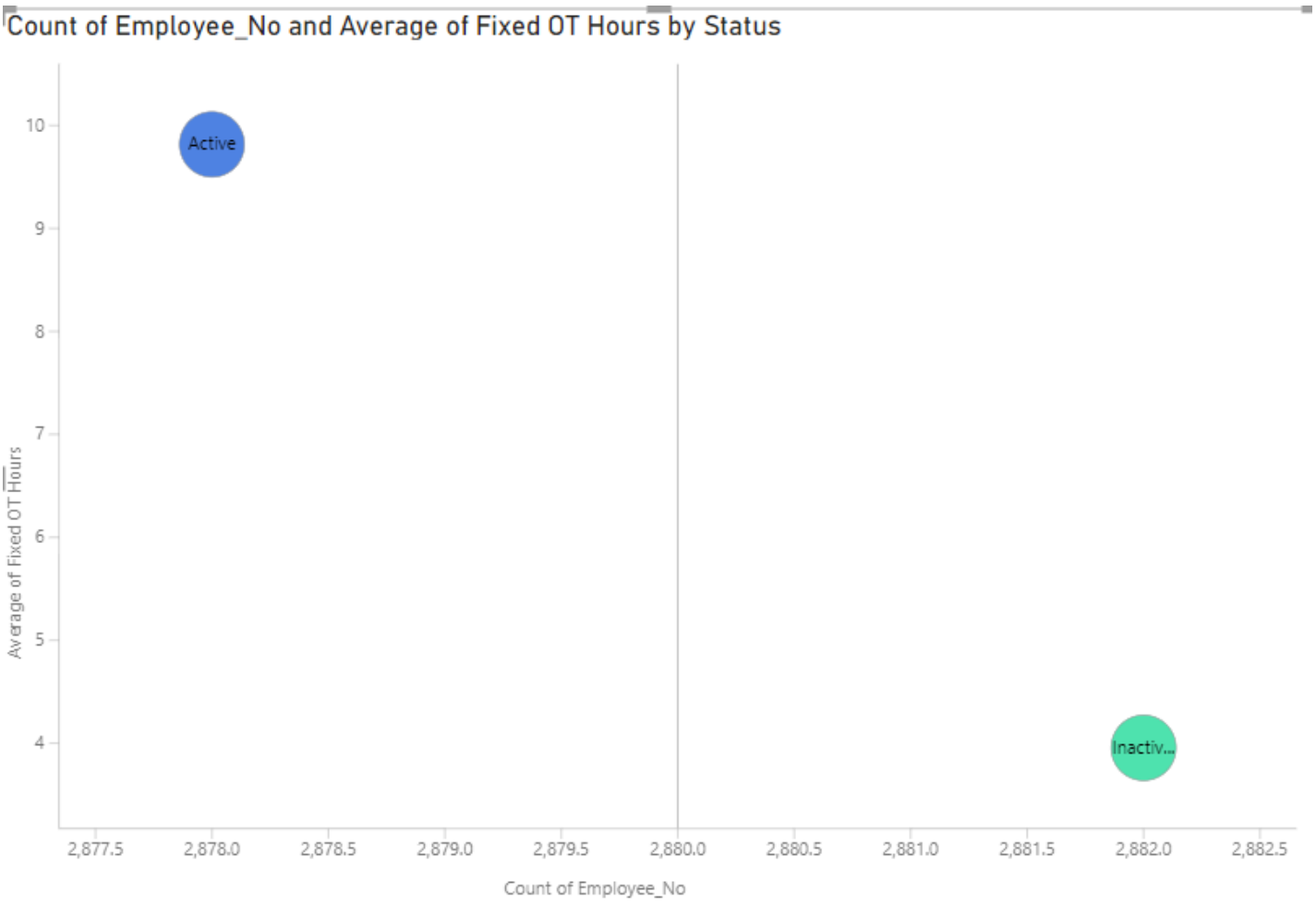


In this scatter plot, the x-axis is representing the employees by their 'Employee_No' and the y-axis is representing the number of attendances of each of them. In here yellow dots mean inactive employees and blue dots represent active employees. By looking at this graph we can see that there is a **negative relationship between attendance and the activeness of the employees**.

After getting **the correlation** between these two labels, the value was **-0.55** and it is considerably highly related. **Active employees tend to have good attendance maintenance and inactive employees had a smaller number of attendances** most of the time.

By considering this relation, the company must consider the attendance of each employee. If the employee has a smaller number of attendance other than the other employees, he tends to leave the company.

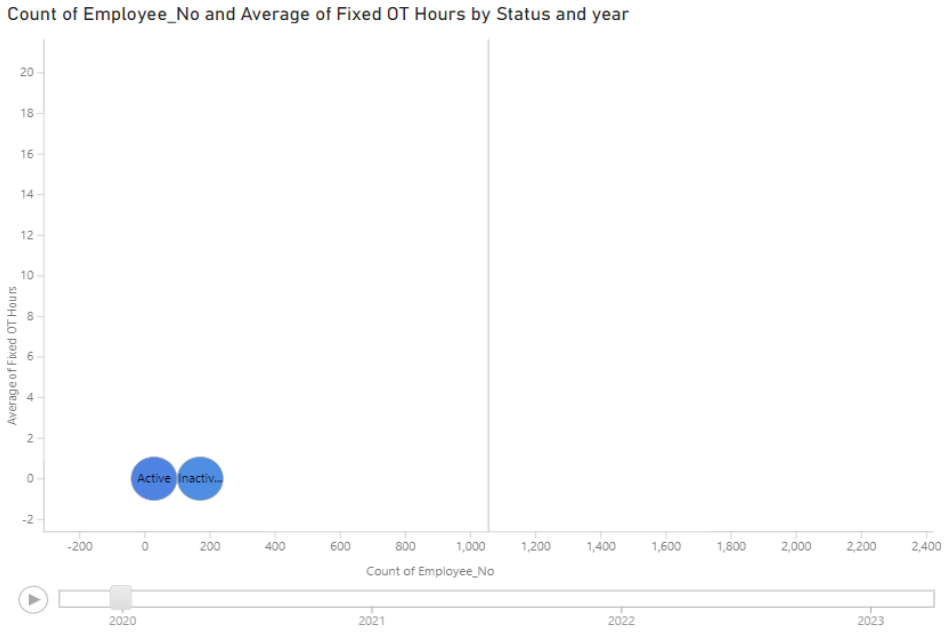
- **COUNT OF EMPLOYEES VS STATUS, FIXED OT HOURS & YEAR**



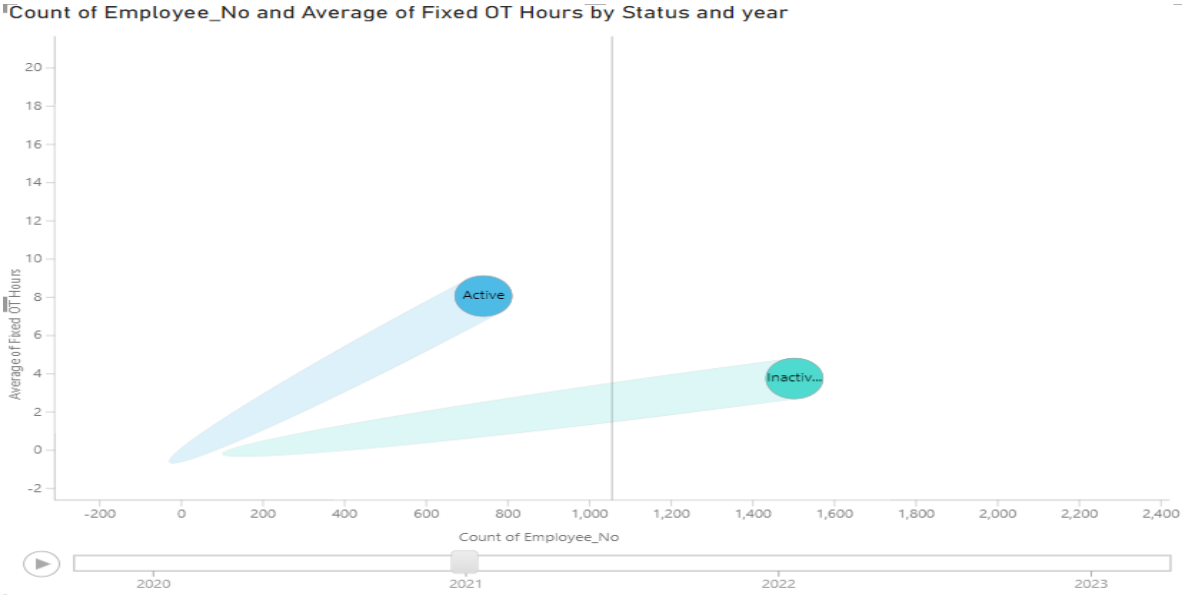
In the given graph, the x-axis represents the count of the employees and the y-axis represents the average fixed OT hours. The green dot represents the inactive employees and the blue dot represents the active employees.

By looking at the graph, we can see that **active employees have a larger number of OT hours other than inactive employees**. The average OT hours of active employees are 9.82 hours and 2878 employees in active employees while the average OT hours of inactive employees are 3.96 hours and 2882 employees in inactive employees. Most of the time in the data set, inactive members have less number of fixed OT hours than active members.

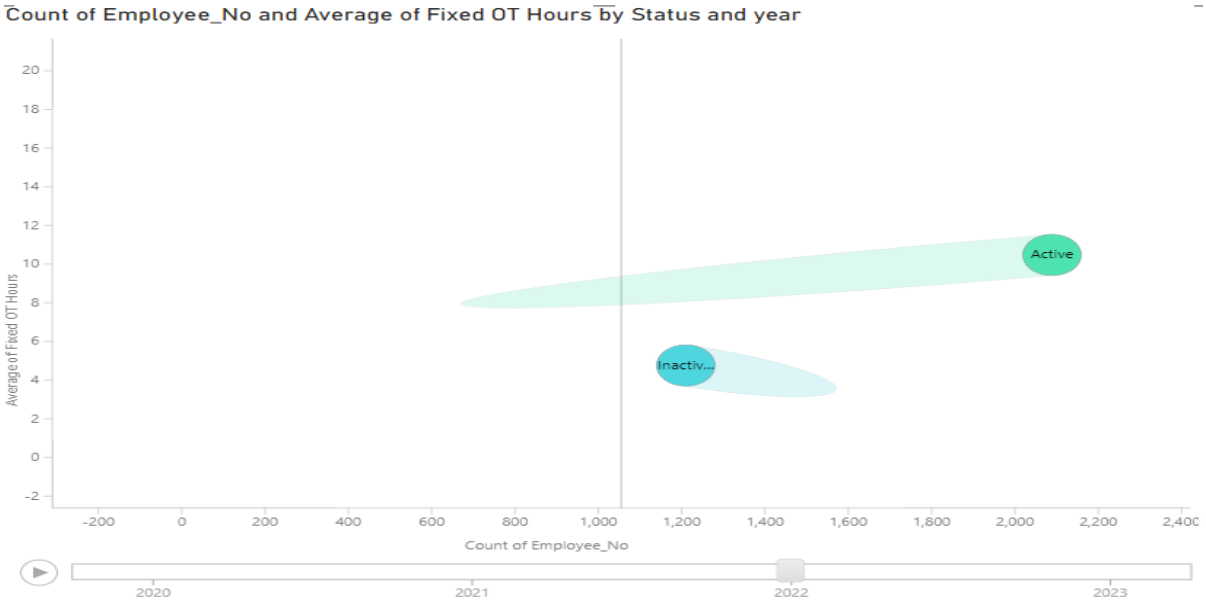
We can see this graph for the years 2020, 2021, 2022, 2023. In those graphs, we can see how the activeness got changed when the year changed. The following graph shows it for 2020.In that graph, there are no Fixed OT hour records in the records. It shows 198 employees consisting of 28 active employees and 170 inactive employees all of whom have 0 fixed OT hours.



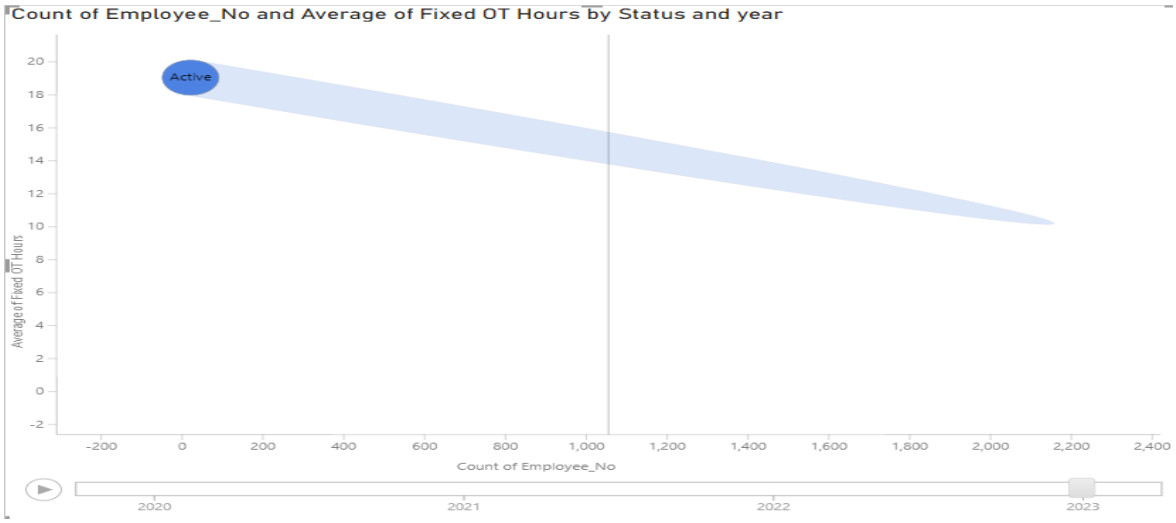
In the 2021 graph, active employees are 740 number, and they have an average 8.07 of fixed OT hours and there are 1502 inactive employees with 3.75 fixed OT hours. By referring to this graph we can say that employees who got more fixed OT hours tend to be in the group and employees who got a low number of fixed OT hours tend to leave the company in the year 2021.



In the year 2022, active employees are higher than inactive employees When considering the fixed OT hours, we can see that the employees who got higher OT hours tend to stay in the company and employees who got small number of hours in fixed OT hours tends to leave the company.



In the 2023 graph, there are only active employees who have more than 18 hours of fixed OT hours on average. By this, we can say that employees tend to be in the company if they get more fixed OT hours.



Links:

Colab Notebooks:

https://colab.research.google.com/drive/1-zxAu0W63lcF2gf_UuGc38g1B12XPpR#scrollTo=D8MeZQAmhQ_f

https://colab.research.google.com/drive/1-_zxAu0W63lcF2gf_UuGc38g1B12XPpR#scrollTo=D8MeZQAmhQ_f

Power BI :

https://app.powerbi.com/groups/me/reports/65949e8f-11dd-46fd-83c6-2fe33110a10/ReportSection95aa6ffdc45483f700c5?experience=power-bi

***** End*****