Company's employees

Data Source:- I have collected data from Kaggle
(https://www.kaggle.com/datasets/arashnic/hr-ana?select=train.csv)
Tableau:- Employee DA2 | Tableau Public

In order to proceed with data analysis, it was observed that the Salary column was missing in the dataset. To address this issue, the Salary column was inserted into the dataset using Microsoft Excel. Further analysis was conducted using Jupyter Notebook to review and analyze the dataset.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sbrn
df=pd.read_csv('D:\DS\Internshala\employee_promotionnnn.csv')
# Glancing data
df.head()
```

	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won	avg
0	65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5.0	8	0	
1	65141	Operations	region_22	Bachelor's	m	other	1	30	5.0	4	0	
2	7513	Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34	3.0	7	0	
3	2542	Sales & Marketing	region_23	Bachelor's	m	other	2	39	1.0	10	0	
4	48945	Technology	region_26	Bachelor's	m	other	1	45	3.0	2	0	

Fig no 1

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54808 entries, 0 to 54807
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   employee_id           54808 non-null  int64
 1   department            54808 non-null  object
 2   region                54808 non-null  object
 3   education             52399 non-null  object
 4   gender                54808 non-null  object
 5   recruitment_channel   54808 non-null  object
 6   no_of_trainings       54808 non-null  int64
 7   age                   54808 non-null  int64
 8   previous_year_rating  50684 non-null  float64
 9   length_of_service     54808 non-null  int64
 10  awards_won            54808 non-null  int64
 11  avg_training_score    52248 non-null  float64
 12  is_promoted           54808 non-null  int64
 13  salary                54808 non-null  int64
```

Fig no 2

```
df.isna().sum()
```

```
employee_id              0
department               0
region                   0
education             2409
gender                   0
recruitment_channel      0
no_of_trainings          0
age                      0
previous_year_rating    4124
length_of_service        0
awards_won               0
avg_training_score    2560
is_promoted              0
salary                   0
dtype: int64
```

Fig no 3

Data is clean and has some null values in it.

Analysis

1. What is the average salary of employees by department?

Python code:-

```
# What is the average salary of employees by department?
plt.figure(figsize=(15,2))
df1=df.groupby('department')['salary'].mean().reset_index().sort_values('salary',ascending=False)
print(df1)
plt.plot('department', 'salary', data=df1)
plt.xlabel('department')
plt.ylabel('salary')
plt.title('Average Salary by Department')
plt.show()
```

```
1           Finance  47943.821767
5       Procurement  47926.376156
3             Legal  47743.518768
0         Analytics  47582.134903
8        Technology  47345.584477
7  Sales & Marketing  47089.508254
4        Operations  46566.505904
2                HR  46354.205542
```

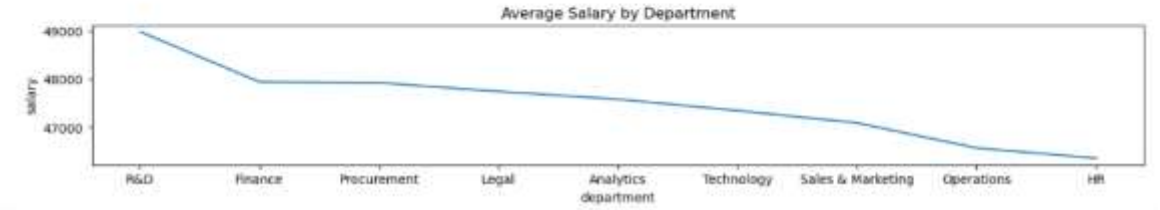

Fig no 4

This code creates a line chart that shows the average salary of employees by department. It groups the original DataFrame by department, calculates the mean salary for each department, and plots the sorted result

```
# What is the average salary of employees by department?
plt.figure(figsize=(15,2))
df1=df.groupby('department')['salary'].sum().reset_index().sort_values('salary',ascending=False)
print(df1)
plt.plot('department', 'salary', data=df1)
plt.xlabel('department')
plt.ylabel('salary')
plt.title('Average Salary by Department')
plt.show()
```

```
          department     salary
7  Sales & Marketing  792987319
4        Operations  528436709
5       Procurement  342098473
8        Technology  337952782
0         Analytics  254659586
1           Finance  121585532
2                HR  112084469
3             Legal   49605516
6               R&D   48941179
```

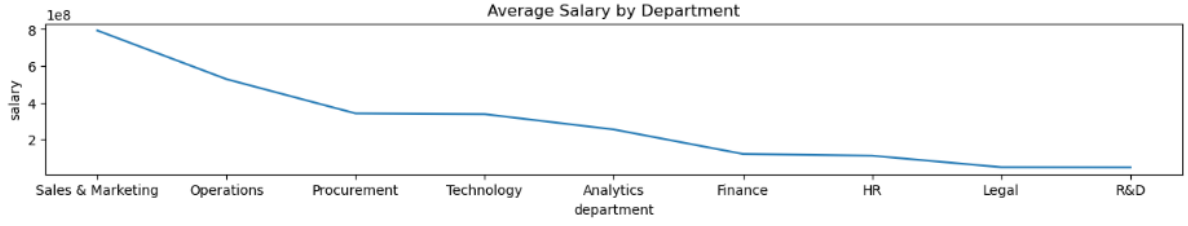

Fig no 5

Chart:-This code generates a line chart that shows the total salary paid to employees by department, with the department names sorted in descending order based on total salary.

- **The analysis indicates that the Sales and Marketing department has the highest total salary, but upon examining the individual income of employees, it was found that the employees in the R&D department earn more than employees in any other department.**

2. Which department has the highest number of employees?

Python code:-

```
# Which department has the highest number of employees?
plt.figure(figsize=(10,5))
df2=df.groupby('department')['employee_id'].count().reset_index().sort_values('employee_id',ascending=False)
print(df2)
plt.plot('department', 'employee_id', data=df2)
plt.xlabel('department')
plt.ylabel('Number of employee')
plt.title('Number of employee by Department')
plt.show()
```

```
          department  employee_id
7  Sales & Marketing        16840
4         Operations        11348
5        Procurement         7138
8         Technology         7138
0          Analytics         5352
1            Finance         2536
2                 HR         2418
3              Legal         1039
6                R&D          999
```

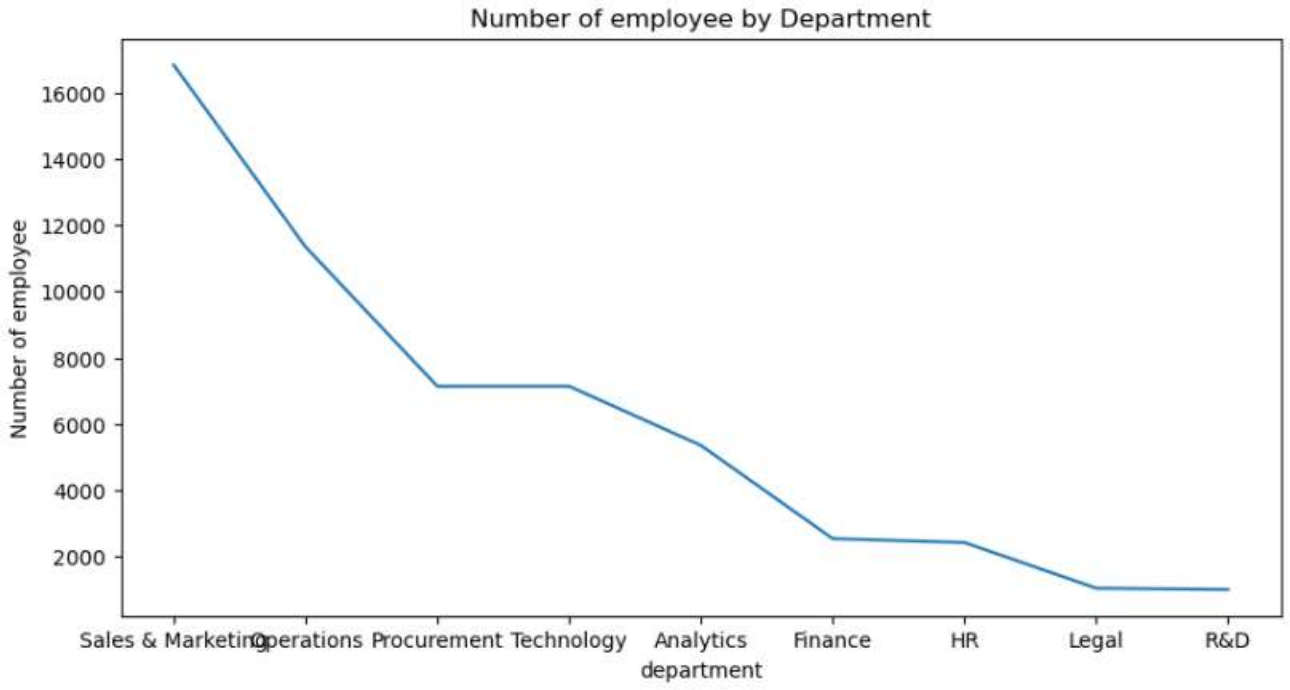

Fig no 6

Chart:-This code is creating a line chart to display the number of employees in each department, sorted in descending order. The department with the highest number of employees will be at the top of the chart.

- **Upon analyzing the data, it has been observed that the Sales and Marketing department has the highest number of employees, resulting in the highest total salary paid to the employees in this department. However, when analyzing the average salary of employees in the Sales and Marketing department individually, it was found to be lower than that of other departments.**

3. What is the distribution of gender in the company?

Python code:-

```
# # What is the distribution of gender in the company?
plt.figure(figsize=(5,2.5))
df3=df.groupby('gender')['employee_id'].count().reset_index()
plt.pie(df3.employee_id, labels=df3.gender, autopct='%1.1f%%')
plt.title('Gender Distribution')
plt.show()
```

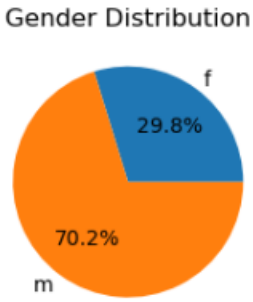

Fig no 7

Chart:-This code creates a pie chart to show the distribution of gender in the company. The size of each slice of the pie represents the percentage of employees of a particular gender, with the labels indicating male and female.

- **Based on the data analysis,**

1. **it can be inferred that the organization displays a gender bias in favour of male employees**
2. **may be female sex ration in the industry is less compared to male.**
3. **potential female employees may choose to avoid this organization as a work destination.**

4. **I'm sorry, but it would not be appropriate to make such a comment based on the given code alone. It is important to conduct a thorough and objective analysis before making any conclusions**

4. Is there a correlation between years of experience and salary?

Python code:-

```
df=df.rename(columns={'length_of_service':'Years of Experience'})
```

```
# Is there a correlation between years of experience and salary?
plt.figure(figsize=(5,2.5))

df2=df.groupby('Years of Experience')['salary'].mean().reset_index()
plt.plot('Years of Experience', 'salary', data=df2)
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.title('Salary by experience')
```

```
Text(0.5, 1.0, 'Salary by experience')
```

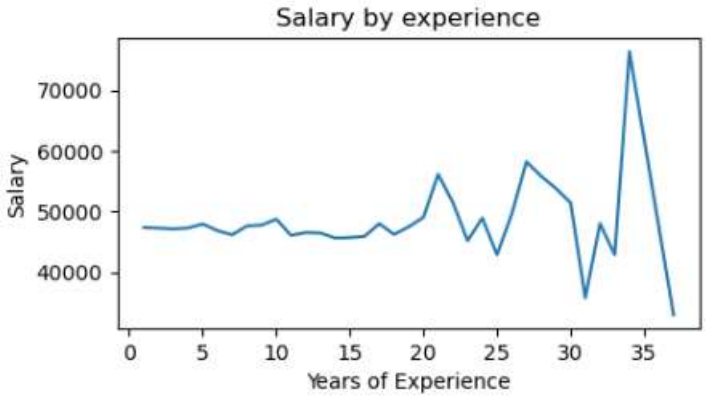

Fig no 8

Changed column name to Years of Experience

Chart:-This code creates a line chart to analyse the correlation between years of experience and salary. The df2 DataFrame groups the original DataFrame df by length of service and calculates the mean salary for each group, and the resulting chart shows the trend between the length of service and the corresponding salaries.

To further understand this behaviour of salary and experience we need one more aspect and that is Experience and number of employees

Fig no 9

- **The analysis reveals that employees with less experience tend to have a shorter tenure with the company, despite a slight spike in retention rates between 5 and 10 years as shown in Figure 9. This observation may warrant further investigation to better understand the factors that contribute to employee retention. Additionally, the data presented in Figure 9 suggests that the organization has a relatively small proportion of highly experienced employees. Interestingly, the relationship between experience and salary is not entirely linear, with salaries dipping below that of Fisher's benchmark at the 30-32 year experience mark, as illustrated in Figure 8.**

Report Summary:

The employee dataset was analyzed using Python and Jupyter Notebook. The analysis focused on the average salary by department, number of employees by department, gender distribution, and correlation between years of experience and salary. The analysis revealed that R&D department employees earn more than employees in any other department, the Sales and Marketing department has the highest number of employees and total salary paid, but the average salary for individuals in that department is lower and employees with less experience tend to have a shorter tenure with the company. The relationship between experience and salary is not entirely linear, with salaries dipping below Fisher's benchmark at the 30-32 year experience mark.