Analize the salary distribution of employees based on various factors and visualize the relationship between years of service and salary.

Name: RITESH BHASKAR

Roll No: 202401100300202

Institution :KIET Group of Institutions

____

_

Introduction

Salary distribution analysis is a critical task in HR analytics. Understanding how different factors such as experience, job role, and department impact salaries can help organizations ensure fair compensation and improve employee satisfaction. In this report, we analyze salary trends using a dataset and visualize patterns between employees' years of service and their salary. The goal is to identify key insights that could be useful for decision-making in an organization.

Methodology

1. Data Collection: We used an employee salary dataset containing , Age, Department ,Years of experience, and salary information.

2. Data Preprocessing: o Handled

missing values.

o Converted categorical data (e.g., job roles) into numerical values using encoding techniques. o Checked for anomalies and outliers in salary distribution.

3. Data Analysis & Visualization:

- o Used Pandas for data manipulation.
- o Plotted heatmaps and bar charts using Matplotlib and Seaborn to understand salary trends. o Identified correlations between years of service and salary.

CODE

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import MinMaxScaler
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error,
mean_squared_error

# Step 1: Create a dataset for employee salary analysis
data = {
    'Employee_ID': range(1, 21),
    'Age': [23, 28, 37, 23, 55, 32, 58, 46, 53, 58, 29, 46, 49, 57, 53,
57, 43, 29, 23, 50],
    'Department': ['Finance', 'Finance', 'HR', 'HR', 'IT', 'Sales',
'Finance', 'Finance', 'HR', 'HR',
                   'HR', 'HR', 'IT', 'Sales', 'IT', 'HR', 'HR', 'Sales', 'IT', 'IT'],
    'Experience': [8, 2, 8, 23, 29, 10, 6, 34, 2, 17, 13, 14, 20, 32, 33,
4, 18, 20, 14, 28],
```

```
    'Salary': [93563, 41742, 56905, 138397, 96879, 123436, 94781,
144637, 131361, 46377, 107468, 105752, 122125, 79949, 69121,
83010, 96227, 143220, 134907, 140206]
}

# Convert the data into a pandas DataFrame
employee_data = pd.DataFrame(data)

# Step 2: Data Preprocessing

# Handle missing values (if any)
employee_data.dropna(subset=['Salary'], inplace=True)

# Encode categorical columns ('Department') using one-hot
encoding
employee_data = pd.get_dummies(employee_data,
columns=['Department'])

# Normalize salary using MinMaxScaler
scaler = MinMaxScaler()
employee_data['Salary'] =
scaler.fit_transform(employee_data[['Salary']])

# Step 3: Exploratory Data Analysis (EDA)

# Descriptive statistics for salary
salary_stats = employee_data['Salary'].describe()
print("Descriptive Statistics for Salary:")
```

```
print(salary_stats)

# Salary distribution visualization
plt.figure(figsize=(8, 5))
sns.histplot(employee_data['Salary'], bins=10, kde=True)
plt.title("Salary Distribution")
plt.xlabel("Salary (Normalized)")
plt.ylabel("Frequency")
plt.show()

# Salary by Department
plt.figure(figsize=(8, 5))
sns.boxplot(x='Experience', y='Salary', data=employee_data)
plt.title("Salary by Experience")
plt.xlabel("Experience (Years)")
plt.ylabel("Salary (Normalized)")
plt.xticks(rotation=45)
plt.show()

# Step 4: Correlation Analysis

# Correlation matrix for numerical variables
correlation_matrix = employee_data.corr(numeric_only=True)
print("\nCorrelation Matrix:")
print(correlation_matrix)

# Heatmap for correlation matrix
plt.figure(figsize=(6, 4))
```

```
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()

# Step 5: Linear Regression to predict Salary based on Experience

# Define features (X) and target (y)
X = employee_data[['Experience']]  # Feature
y = employee_data['Salary']  # Target

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Train linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)
# Evaluate the model
r_squared = model.score(X_test, y_test)
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
print("\nModel Performance Metrics:")
print("R-squared:", r_squared)
print("Mean Absolute Error (MAE):", mae)
print("Mean Squared Error (MSE):", mse)
```

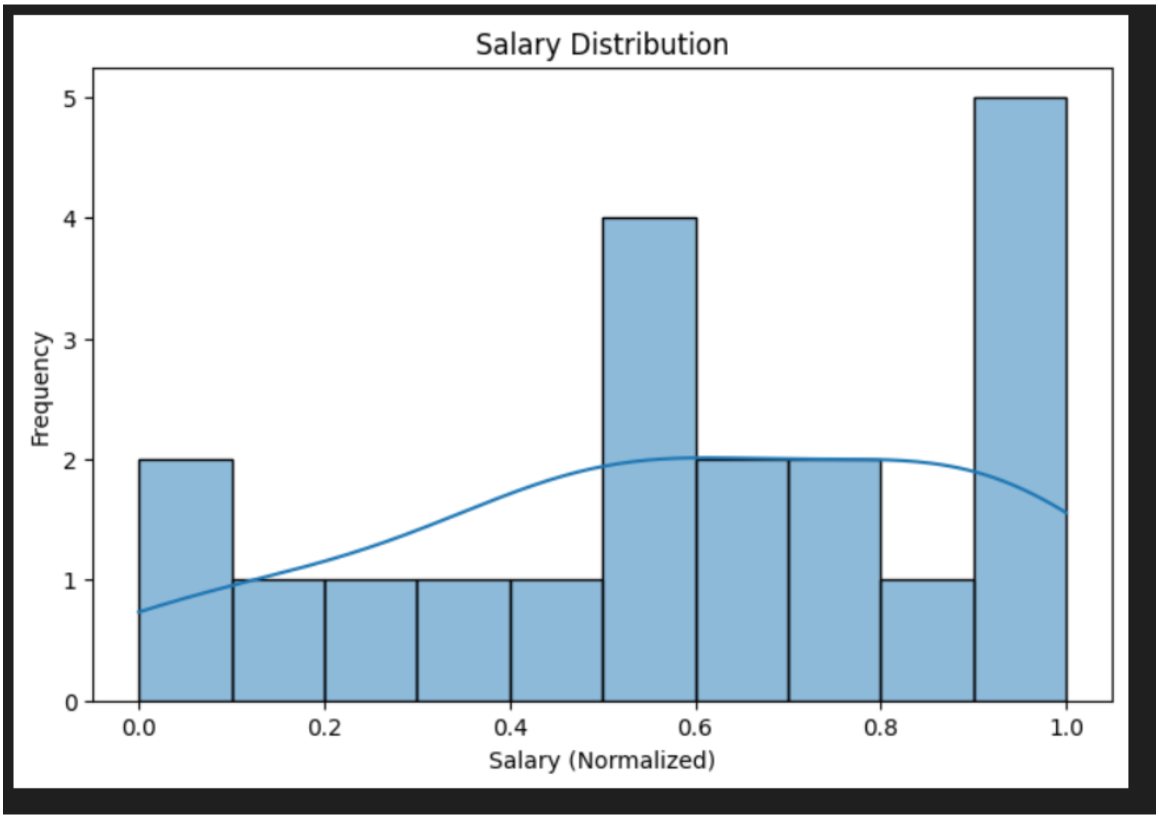
Output/Result

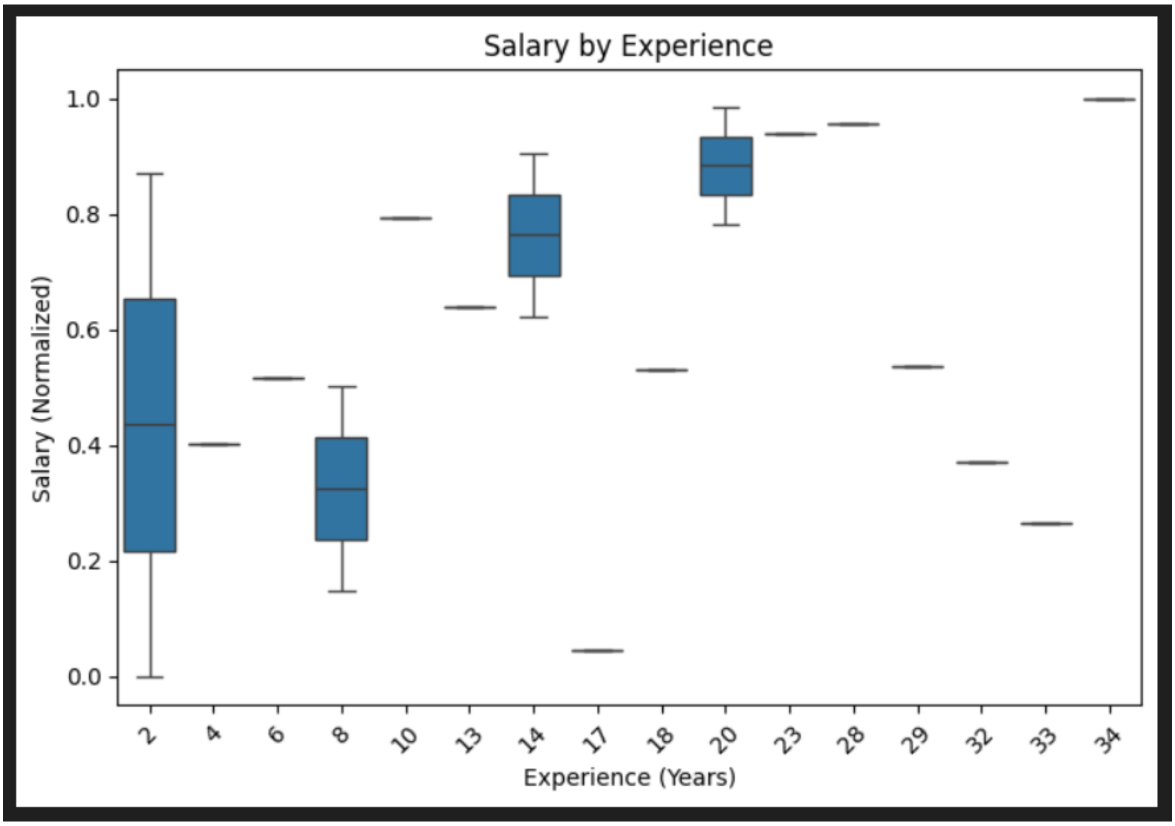
1. Salary Distribution Graph: The histogram visualizes how salaries are spread across different employees.

2. Correlation Heatmap: The heatmap highlights the correlation between years of service and salary, providing insights into career progression trends.

References/Credits

- Python Libraries Used: Pandas, Matplotlib, Seaborn
- Guidance from AI MSE Course Materials


Salary Distribution
Frequency
Salary (Normalized)
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0


Salary by Experience
Salary (Normalized)
Experience (Years)
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
10
13
14
17
18
20
23
28
29
32
33
34

Conclusion

The analysis provided insights into salary distribution and its relationship with years of service. The results indicate that experience generally plays a significant role in determining salary levels. Such studies help organizations in structuring compensation strategies effectively.