AI MSE PROJECT REPORT

Name : Arushi Sagar
Branch : CSE-AIML
Section : A
Roll number : 52
Library ID : 2428CSEAIML2022

Problem Statement : Employee Salary Analysis, explore correlations in employee salaries and positions with visualizations.

INTRODUCTION

This problem focuses on analyzing employee salaries based on different factors such as job position, experience, education, gender and department. By using data visualization techniques, we will explore patterns, trends, and correlations to understand salary distribution and influencing factors.

Methodology Used in the Code:

1. **Data Generation:**
   - Created a synthetic dataset with 200 employees.
   - Assigned attributes like Employee ID, Name, Age, Gender, Department, Position, Years of Experience, and Education Level.
   - Defined a salary range for each position and randomly assigned salaries.
2. **Data Storage:**
   - Converted the generated data into a pandas DataFrame.
   - Saved the dataset as a CSV file for further analysis.
3. **Exploratory Data Analysis:**
   - Used `df.describe()` to compute summary statistics (mean, min, max, etc.).
   - Checked for missing values using `df.isnull().sum()`.
4. **Data Visualization:**
   - **Boxplots:** Analyzed salary distribution across positions, education levels, and departments.
   - **Scatterplot:** Explored the relationship between years of experience and salary.
   - **Violin Plot:** Compared salary distributions across genders.
   - **Heatmaps:** Visualized correlations between age, experience, and salary.

CODE :

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns




# Set seed for reproducibility

np.random.seed(42)


# Generate sample data

num_employees = 200

employee_ids = [f'EMP{str(i).zfill(3)}' for i in range(1, num_employees + 1)]

names = [f'Employee_{i}' for i in range(1, num_employees + 1)]

ages = np.random.randint(22, 60, num_employees)

genders = np.random.choice(['Male', 'Female', 'Other'], num_employees, p=[0.5, 0.45, 0.05])

departments = np.random.choice(['HR', 'Finance', 'IT', 'Marketing', 'Sales', 'Operations'], num_employees)

positions = np.random.choice(['Intern', 'Junior', 'Senior', 'Lead', 'Manager', 'Director'], num_employees, p=[0.1, 0.3, 0.3, 0.15, 0.1, 0.05])

years_experience = np.random.randint(0, 35, num_employees)

education_levels = np.random.choice(['High School', 'Bachelor', 'Master', 'PhD'], num_employees, p=[0.2, 0.5, 0.25, 0.05])
```

```python
# Salary distribution based on position
position_salary_map = {
    'Intern': (30000, 40000),
    'Junior': (40000, 60000),
    'Senior': (60000, 90000),
    'Lead': (90000, 120000),
    'Manager': (120000, 150000),
    'Director': (150000, 200000)
}

salaries = [np.random.randint(position_salary_map[pos][0], position_salary_map[pos][1]) for pos in positions]

# Create DataFrame
df = pd.DataFrame({
    'Employee ID': employee_ids,
    'Name': names,
    'Age': ages,
    'Gender': genders,
    'Department': departments,
    'Position': positions,
    'Years of Experience': years_experience,
    'Education Level': education_levels,
    'Salary': salaries
})

# Save to CSV
df.to_csv('employee_salary_dataset.csv', index=False)
```

```
# Display first few rows
df.head()
```

```
print(df.shape) # Shows the dimensions of a DataFrame
```

```
# Summary Statistics
print(df.describe())# Provides statistical summary including mean, standard deviation, min, and max values
```

```
print(df.dtypes) #Shows different data types
```

```
print(df.isnull().sum())   # Prints the count of missing values per column
```

```
# Data Visualization
#Salary distribution by position
```

```
plt.figure(figsize=(10, 6))

sns.boxplot(x='Position', y='Salary', data=df, order=['Intern', 'Junior', 'Senior', 'Lead', 'Manager', 'Director'])

plt.title('Salary Distribution by Position')

plt.xticks(rotation=45)

plt.show()
```

```
#Salary vs years of experience data visualization

plt.figure(figsize=(10, 6))

sns.scatterplot(x='Years of Experience', y='Salary', hue='Position', data=df)

plt.title('Salary vs. Years of Experience')

plt.show()
```

```
#Salary distribution by department

plt.figure(figsize=(10, 6))

sns.boxplot(x='Department', y='Salary', data=df)

plt.title('Salary Distribution by Department')

plt.xticks(rotation=45)

plt.show()
```

```
#Salary distrubution by gender

plt.figure(figsize=(10, 6))
```

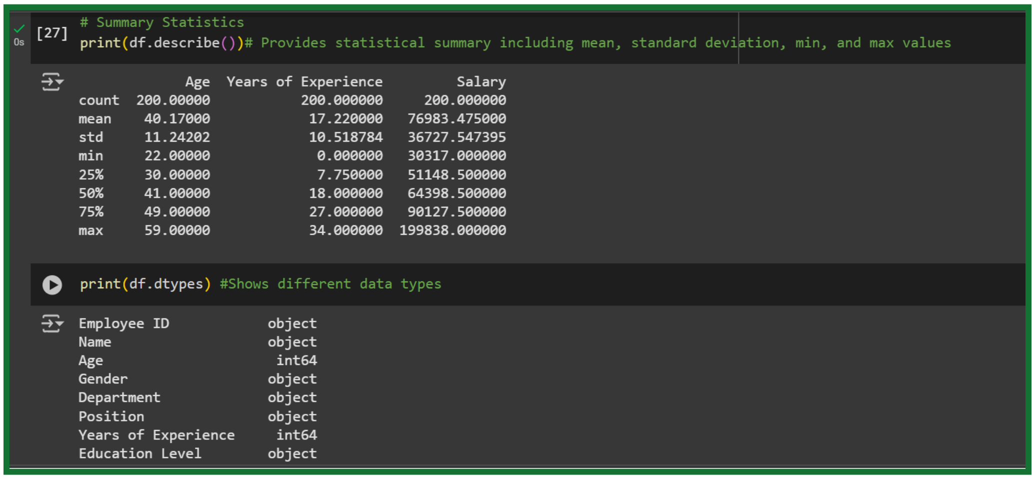
```
sns.violinplot(x='Gender', y='Salary', data=df)

plt.title('Salary Distribution by Gender')

plt.show()
```

OUTPUT :

output of describe,data type and isnull command :

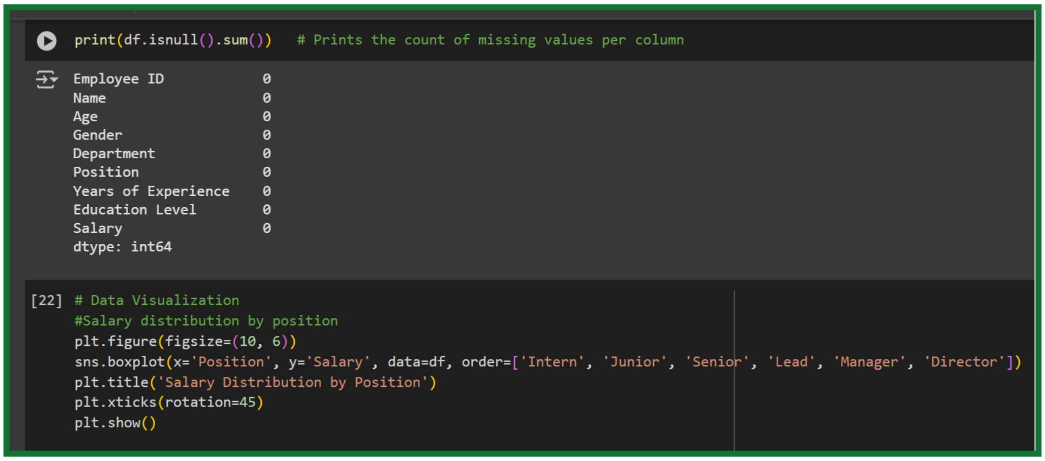


```
[27] # Summary Statistics
     print(df.describe())# Provides statistical summary including mean, standard deviation, min, and max values

                Age  Years of Experience         Salary
count  200.00000           200.000000     200.000000
mean    40.17000            17.220000   76983.475000
std     11.24202            10.518784   36727.547395
min     22.00000             0.000000   30317.000000
25%     30.00000             7.750000   51148.500000
50%     41.00000            18.000000   64398.500000
75%     49.00000            27.000000   90127.500000
max     59.00000            34.000000  199838.000000
```

```
print(df.dtypes) #Shows different data types

Employee ID            object
Name                   object
Age                     int64
Gender                 object
Department             object
Position               object
Years of Experience     int64
Education Level        object
```

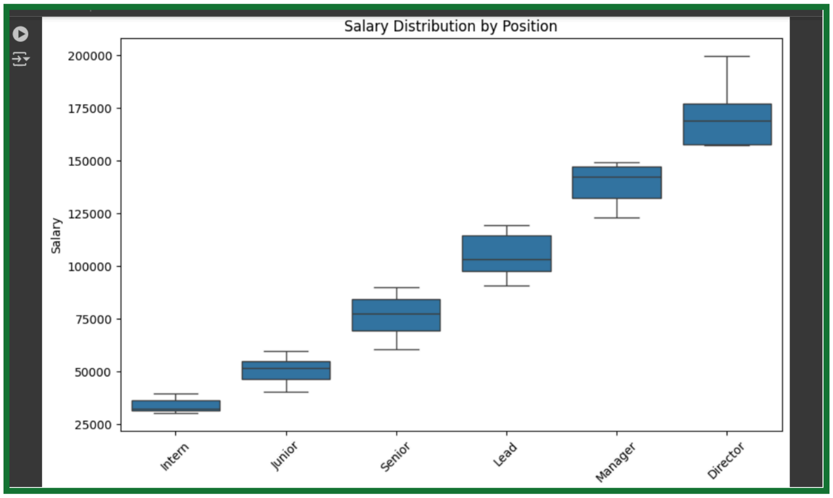


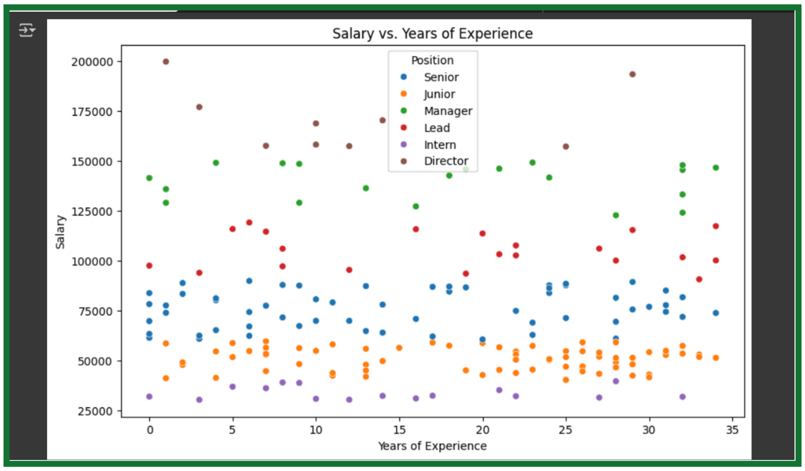
```
print(df.isnull().sum())    # Prints the count of missing values per column

Employee ID            0
Name                   0
Age                    0
Gender                 0
Department             0
Position               0
Years of Experience    0
Education Level        0
Salary                 0
dtype: int64
```

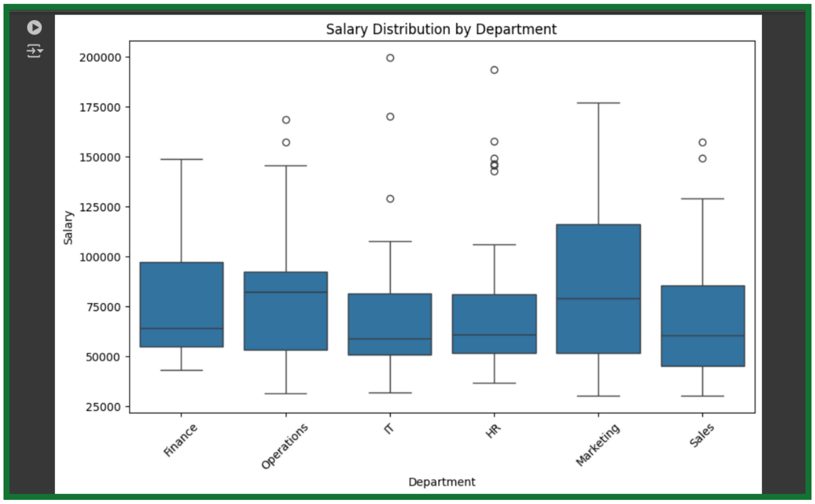
```
[22] # Data Visualization
     #Salary distribution by position
     plt.figure(figsize=(10, 6))
     sns.boxplot(x='Position', y='Salary', data=df, order=['Intern', 'Junior', 'Senior', 'Lead', 'Manager', 'Director'])
     plt.title('Salary Distribution by Position')
     plt.xticks(rotation=45)
     plt.show()
```

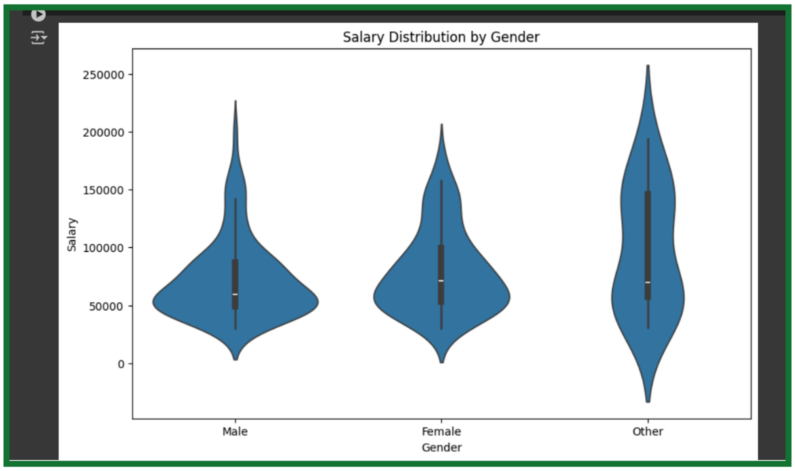
Graphs of salary distribution by position and salary vs years of experience :





Graphs of salary distribution by department and gender :


Salary Distribution by Department
Salary
Finance
Operations
IT
HR
Marketing
Sales
Department


Salary Distribution by Gender
Salary
Male
Female
Other
Gender