# NEXUS SOFTWARE

# DATA ANALYSIS-PROJECT_1

**Name of the project:** Employee Analysis (Entry Level)

**Objective:** Conduct basic data analytics and visualization on the Employee DataSet to gain insights into the workforce. No machine learning is required for this entry-level project, and tools such as PowerBI, Tableau, or any preferred tool can be used for analysis and visualization.

**Tools I used**: Google colab

**Task -1:**

**Descriptive Statistics:**

- Calculate and present the average age, average experience in the current

domain, and average years since joining for the entire workforce.

**Calculating average age**:

## Calculating average age

```
# Calculate descriptive statistics
average_age = df['Age'].mean()
print(f"Average Age: {average_age:.2f} years")

Average Age: 29.39 years
```

**Calculating average experience in current domain :**

## Calculating average experience in current domain

```
average_experience = df['ExperienceInCurrentDomain'].mean()
print(f"Average Experience in Current Domain: {average_experience:.2f} years")

Average Experience in Current Domain: 2.91 years
```

**Average years since joining for the entire workforce:**

```
Average years since joining for the entire workforce.

average_years_since_joining = pd.Timestamp.now().year - df['JoiningYear'].mean()
print(f"Average Years Since Joining: {average_years_since_joining:.2f} years")

Average Years Since Joining: 7.94 years
```

**Task -2:**

City-wise Analysis:

- Explore the distribution of employees across different cities using charts or maps to identify significant locations.

```python
# City-wise distribution for the bar chart
city_distribution = df['City'].value_counts()

#Plotting a bar chart
plt.figure(figsize=(10, 6))
city_distribution.plot(kind='bar', color='skyblue')
plt.title('City-wise Distribution of Employees')
plt.xlabel('City')
plt.ylabel('Number of Employees')
plt.xticks(rotation=45, ha='right')
plt.show()
```
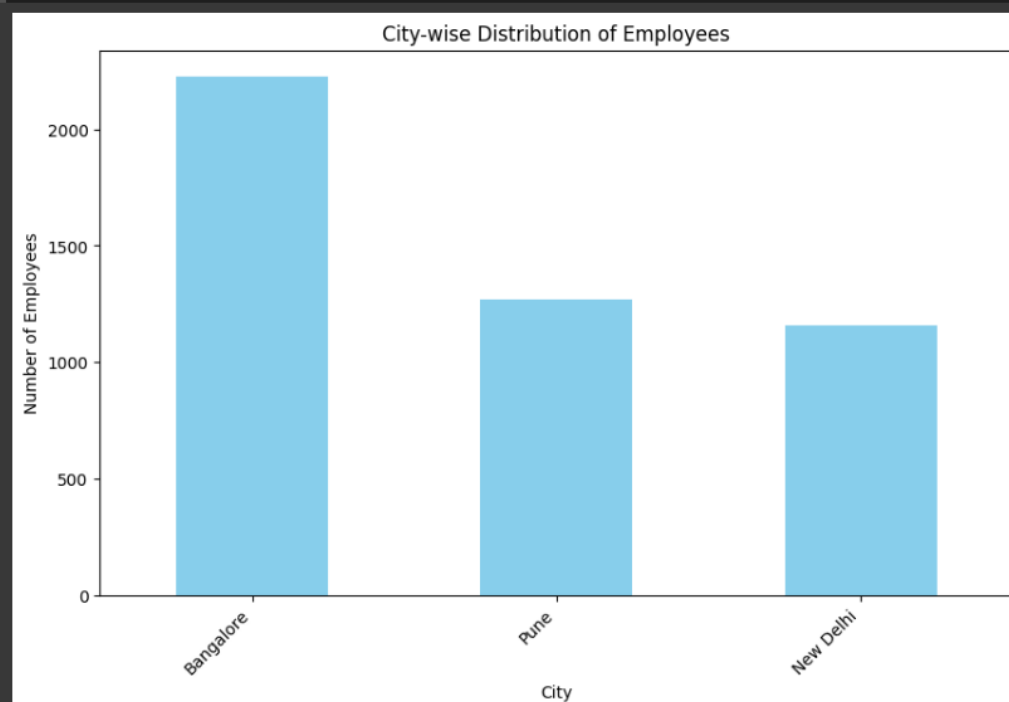
**Task -3:**

**- Visualize the gender distribution within the company to understand the level of gender diversity.**

```python
gender_distribution = df['Gender'].value_counts()

# Plotting a pie chart
plt.figure(figsize=(8, 8))
plt.pie(gender_distribution, labels=gender_distribution.index, autopct='%1.1f%%', colors=['skyblue', 'lightcoral'])
plt.title('Gender Distribution within the Company')
plt.show()
```



Gender Distribution within the Company
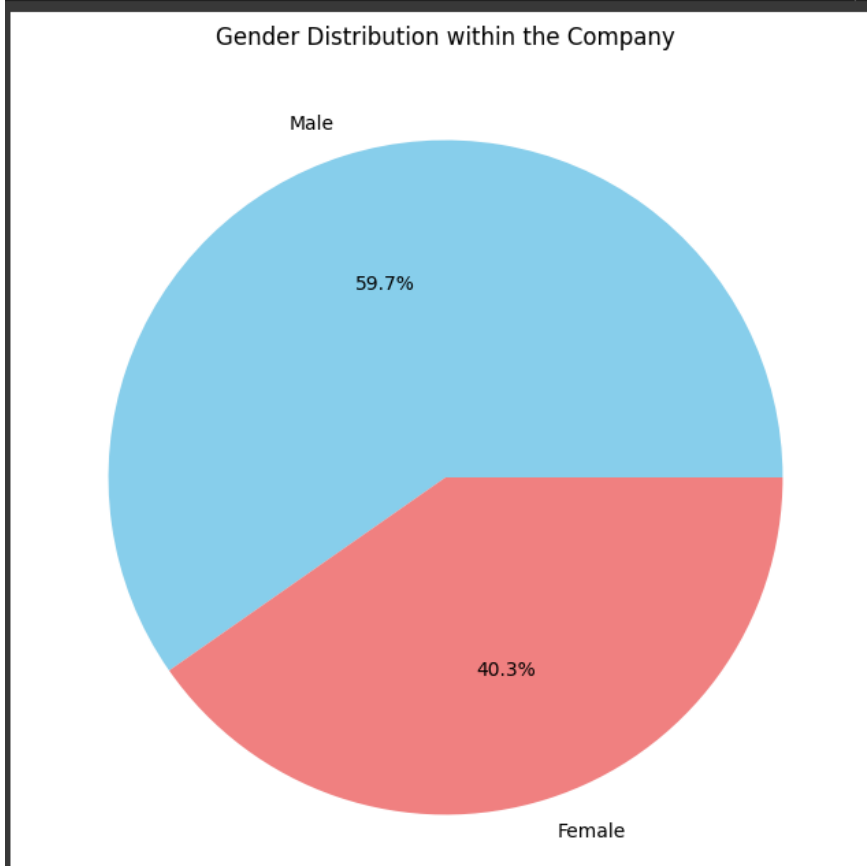
**Task -4:**

- Analyze and present the most common degrees and institutions among

employees.

```python
# Most common degrees for the bar chart
degrees_distribution = df['Education'].value_counts()

# Plotting a bar chart for degrees
plt.figure(figsize=(10, 6))
degrees_distribution.plot(kind='bar', color='skyblue')
plt.title('Most Common Degrees Among Employees')
plt.xlabel('Degree')
plt.ylabel('Number of Employees')
plt.xticks(rotation=0)
plt.show()

# Check if 'Institution' column exists in the dataset
if 'Institution' in df.columns:
    # Most common institutions for the bar chart
    institutions_distribution = df['Institution'].value_counts()

    # Plotting a bar chart for institutions
    plt.figure(figsize=(12, 6))
    institutions_distribution.head(10).plot(kind='bar', color='lightcoral')  # Displaying the top 10 institutions
    plt.title('Most Common Institutions Among Employees')
    plt.xlabel('Institution')
    plt.ylabel('Number of Employees')
    plt.xticks(rotation=45, ha='right')
    plt.show()
else:
    print("No 'Institution' column in the dataset.")
```



No 'Institution' column in the dataset.
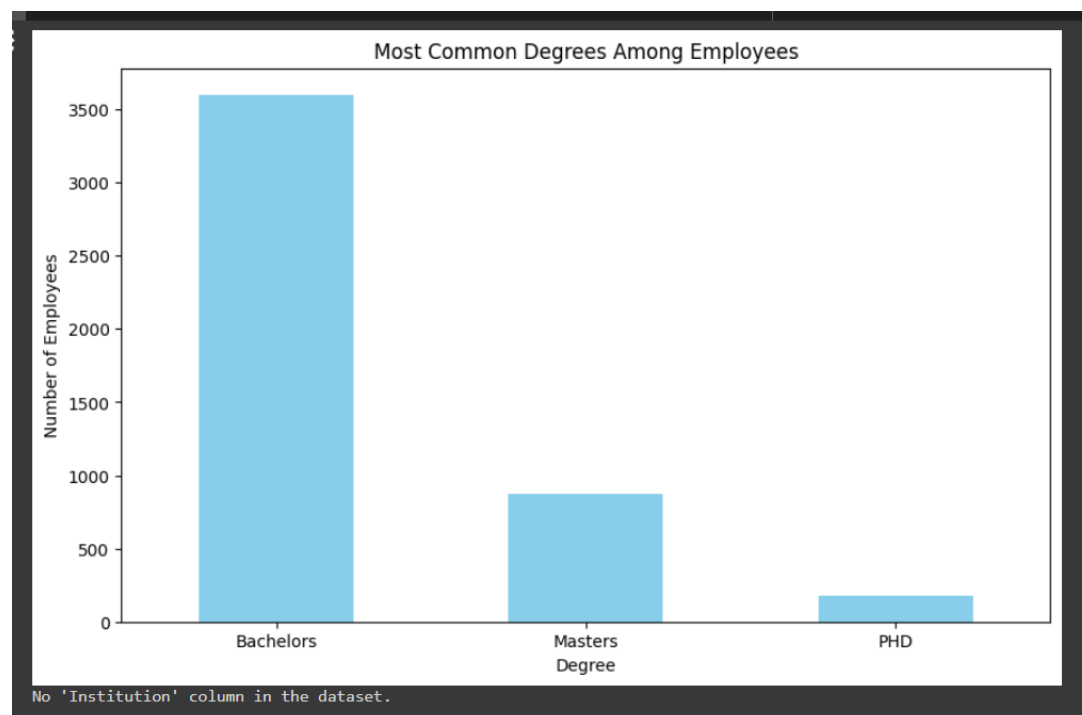
**Task -5:**

- Investigate the distribution of employees across different payment tiers and

visualize the salary distribution to identify trends.

```python
# Payment tier distribution for the bar chart
payment_tier_distribution = df['PaymentTier'].value_counts()

# Plotting a bar chart for payment tier distribution
plt.figure(figsize=(10, 6))
payment_tier_distribution.sort_index().plot(kind='bar', color='skyblue')
plt.title('Distribution of Employees Across Payment Tiers')
plt.xlabel('Payment Tier')
plt.ylabel('Number of Employees')
plt.xticks(rotation=0)
plt.show()
```
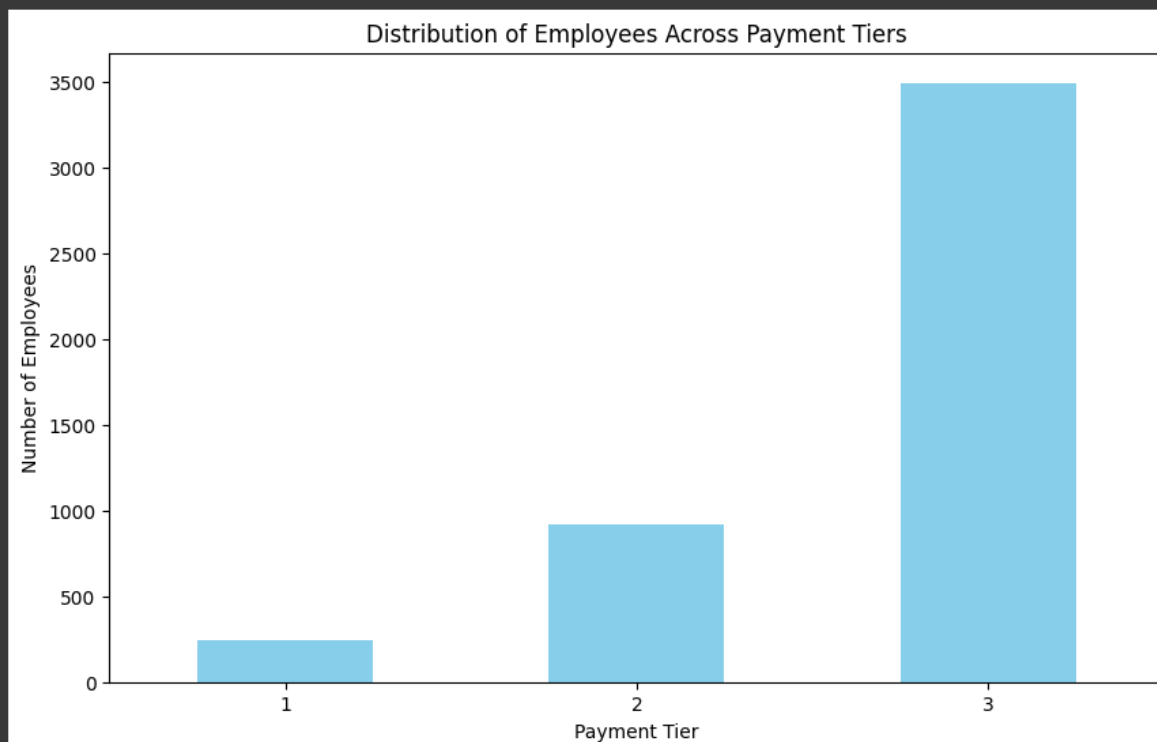
we can visualize the trends in some other numerical column, like 'Age' or 'ExperienceInCurrentDomain' using Histograms:

```python
# Plotting histograms for Age and ExperienceInCurrentDomain
fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(15, 6))

# Histogram for Age
axes[0].hist(df['Age'], bins=20, color='skyblue', edgecolor='black')
axes[0].set_title('Distribution of Age Among Employees')
axes[0].set_xlabel('Age')
axes[0].set_ylabel('Number of Employees')

# Histogram for ExperienceInCurrentDomain
axes[1].hist(df['ExperienceInCurrentDomain'], bins=20, color='lightcoral', edgecolor='black')
axes[1].set_title('Distribution of Experience In Current Domain Among Employees')
axes[1].set_xlabel('Experience In Current Domain')
axes[1].set_ylabel('Number of Employees')

plt.tight_layout()
plt.show()
```
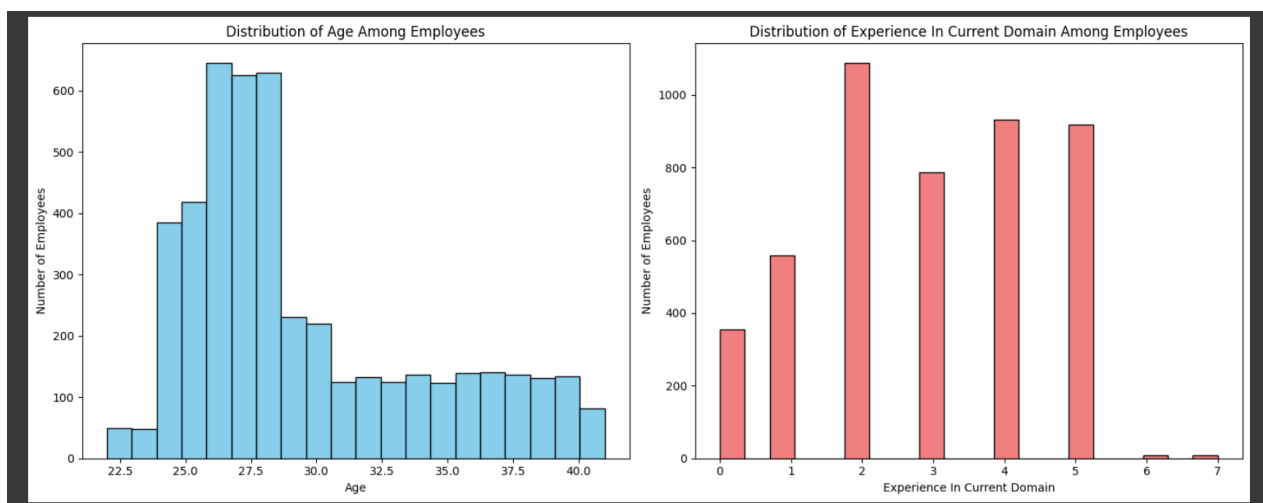
**Task -6:**

- Explore factors influencing leave decisions. Visualize the correlation between factors like age, experience, and the decision to take leave.

```
# Visualizing the correlation matrix as a heatmap
correlation_matrix = df[['Age', 'ExperienceInCurrentDomain', 'LeaveOrNot']].corr()
plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap - Age, Experience, LeaveOrNot')
plt.show()
```



Correlation Heatmap - Age, Experience, LeaveOrNot
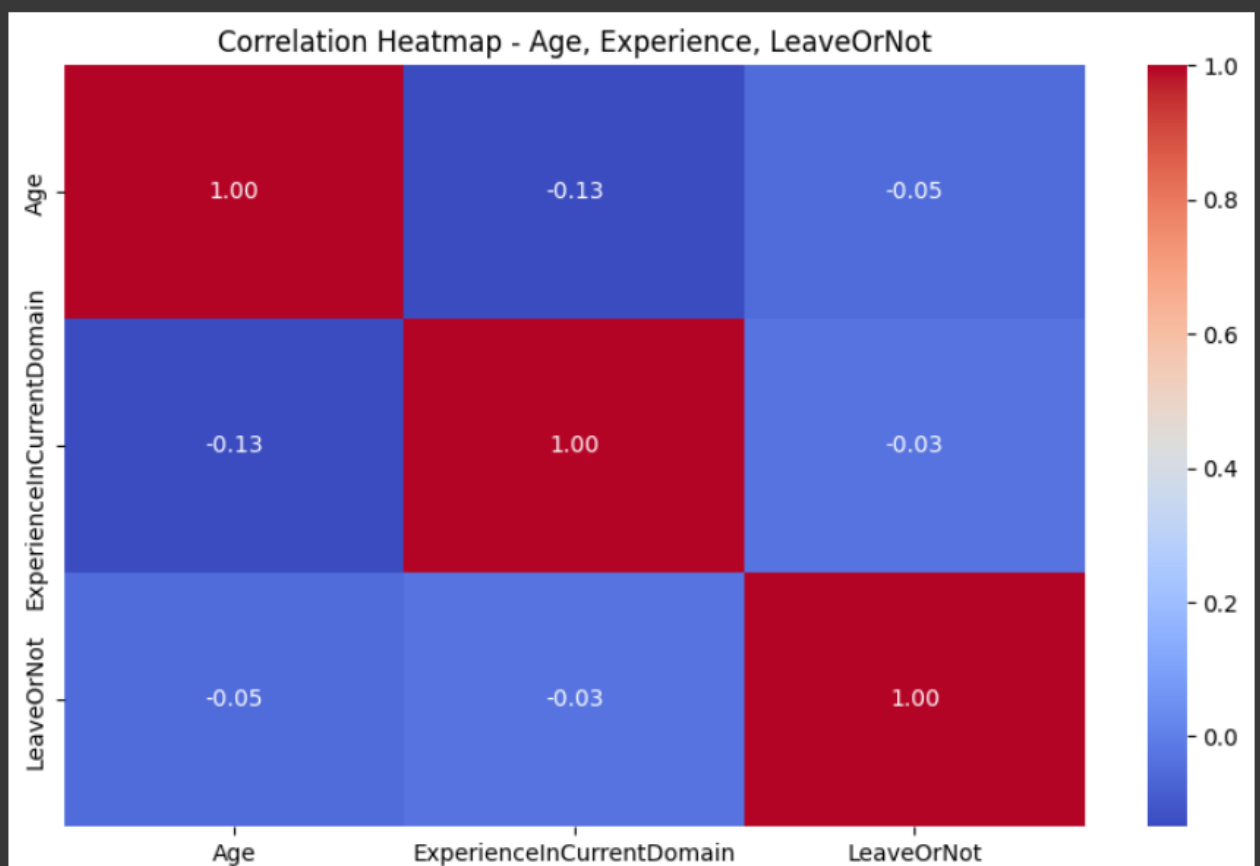
**Summary of the Project:**

## Task -1 :

The code calculates and presents three key descriptive statistics for the entire workforce:

**1. Average Age:**

  - Calculates the mean age of employees in the dataset.

  - Prints the result with two decimal places.

**2. Average Experience in Current Domain:**

  - Calculates the mean number of years of experience employees have in their current field.

  - Prints the result with two decimal places.

**3. Average Years Since Joining:**

  - Calculates the mean number of years since employees joined the company.

  - Prints the result with two decimal places.

The output provides a concise overview of the workforce's age, experience in the current domain, and tenure with the company.

## Task -2 :

Visualizes the distribution of employees across different cities using a bar chart. It calculates the count of employees in each city (city_distribution) and plots a bar chart to show the city-wise distribution. The chart provides insights into the geographical distribution of the workforce.

## Task -3 :

Visualizes the gender distribution within the company using a pie chart. It calculates the count of each gender (`gender_distribution`) and represents the distribution in a pie chart, providing a quick overview of gender diversity within the organization. The chart includes labels and percentages for each gender.

## Task -4 :

The code analyzes and presents the most common degrees and institutions among employees. It includes two visualizations:

**1. Most Common Degrees:**

  - Calculates the count of each degree ('Bachelors', 'Masters', 'PhD').

  - Plots a bar chart to visualize the distribution of degrees among employees.

**2. Most Common Institutions (if 'Institution' column exists):**

- Checks if the 'Institution' column exists in the dataset.

- Calculates the count of employees from each institution.

- Plots a bar chart displaying the top 10 most common institutions among employees.

The visualizations offer insights into the educational background of the workforce, showcasing the prevalence of different degrees and institutions. If the 'Institution' column is not present, a message is printed to inform about its absence in the dataset.

## Task -5:

The code investigates the distribution of employees across different payment tiers and visualizes the salary distribution to identify trends. It includes:

**1. Bar Chart for Payment Tier Distribution:**

- Calculates the count of employees in each payment tier.

- Plots a bar chart to visualize the distribution across payment tiers.

**2. Histograms for Age and Experience in Current Domain:**

- Creates two histograms for age and experience in the current domain.

- Provides insights into the distribution of age and experience among employees.

The visualizations offer a comprehensive view of the salary distribution and additional insights into the age and experience characteristics of the workforce. Adjustments to the code comments may be needed for clarity, as there is a mention of payment tier but histograms for age and experience.

## Task -6 :

The code explores factors influencing leave decisions by visualizing the correlation matrix between age, experience in the current domain, and the decision to take leave (LeaveOrNot). The key elements include:

**Correlation Heatmap:**

Constructs a correlation matrix for age, experience, and leave decisions.

Utilizes a heatmap to visualize the correlations, with annotations and a coolwarm color map.

Provides a concise overview of the relationships between age, experience, and leave decisions.

The heatmap aids in identifying potential correlations between these factors, offering insights into how age and experience may influence the likelihood of employees taking leave. The visualization provides a clear representation of the interplay between variables.