Import Pandas

In [15]:

```
import pandas as pd
df = pd.read_csv("Employee.csv")
df
```

Out[15]:

	Emp_id	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	LeaveOrN
0	E001	Bachelors	2017	Bangalore	3	34	Male	No	0	
1	E002	Bachelors	2013	Pune	1	28	Female	No	3	
2	E003	Bachelors	2014	New Delhi	3	38	Female	No	2	
3	E004	Masters	2016	Bangalore	3	27	Male	No	5	
4	E005	Masters	2017	Pune	3	24	Male	Yes	2	
...	...	...	...	...	...	...	...	...	...	
4648	E4649	Bachelors	2013	Bangalore	3	26	Female	No	4	
4649	E4650	Masters	2013	Pune	2	37	Male	No	2	
4650	E4651	Masters	2018	New Delhi	3	27	Male	No	5	
4651	E4652	Bachelors	2012	Bangalore	3	30	Male	Yes	2	
4652	E4653	Bachelors	2015	Bangalore	3	33	Male	Yes	4	

4653 rows × 10 columns

In [5]:

```
df.sample(5)
```

Out[5]:

	Emp_id	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	LeaveOrN
1658	E1659	Bachelors	2016	Bangalore	3	26	Male	No	4	
29	E030	Masters	2017	New Delhi	2	30	Female	No	2	
1570	E1571	Bachelors	2013	Bangalore	1	24	Male	No	2	
1332	E1333	Bachelors	2012	Pune	3	25	Female	No	3	
1275	E1276	Masters	2015	New Delhi	2	26	Female	No	4	

In [7]:

```
df.describe()
```

Out[7]:

	JoiningYear	PaymentTier	Age	ExperienceInCurrentDomain	LeaveOrNot
count	4653.000000	4653.000000	4653.000000	4653.000000	4653.000000
mean	2015.062970	2.698259	29.393295	2.905652	0.343864
std	1.863377	0.561435	4.826087	1.558240	0.475047
min	2012.000000	1.000000	22.000000	0.000000	0.000000
25%	2013.000000	3.000000	26.000000	2.000000	0.000000
50%	2015.000000	3.000000	28.000000	3.000000	0.000000
75%	2017.000000	3.000000	32.000000	4.000000	1.000000
max	2018.000000	3.000000	41.000000	7.000000	1.000000

In [8]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4653 entries, 0 to 4652
Data columns (total 10 columns):
 #   Column                     Non-Null Count  Dtype 
---  ------                     --------------  ----- 
 0   Emp_id                     4653 non-null   object
 1   Education                  4653 non-null   object
 2   JoiningYear                4653 non-null   int64 
 3   City                       4653 non-null   object
 4   PaymentTier                4653 non-null   int64 
 5   Age                        4653 non-null   int64 
 6   Gender                     4653 non-null   object
 7   EverBenched                4653 non-null   object
 8   ExperienceInCurrentDomain  4653 non-null   int64 
 9   LeaveOrNot                 4653 non-null   int64 
dtypes: int64(5), object(5)
memory usage: 363.6+ KB
```

```
In [10]: df.isnull().sum()
```

```
Out[10]: Emp_id                       0
         Education                    0
         JoiningYear                  0
         City                         0
         PaymentTier                  0
         Age                          0
         Gender                       0
         EverBenched                  0
         ExperienceInCurrentDomain    0
         LeaveOrNot                   0
         dtype: int64
```

```
In [12]: df.duplicated().sum()
```

```
Out[12]: np.int64(0)
```

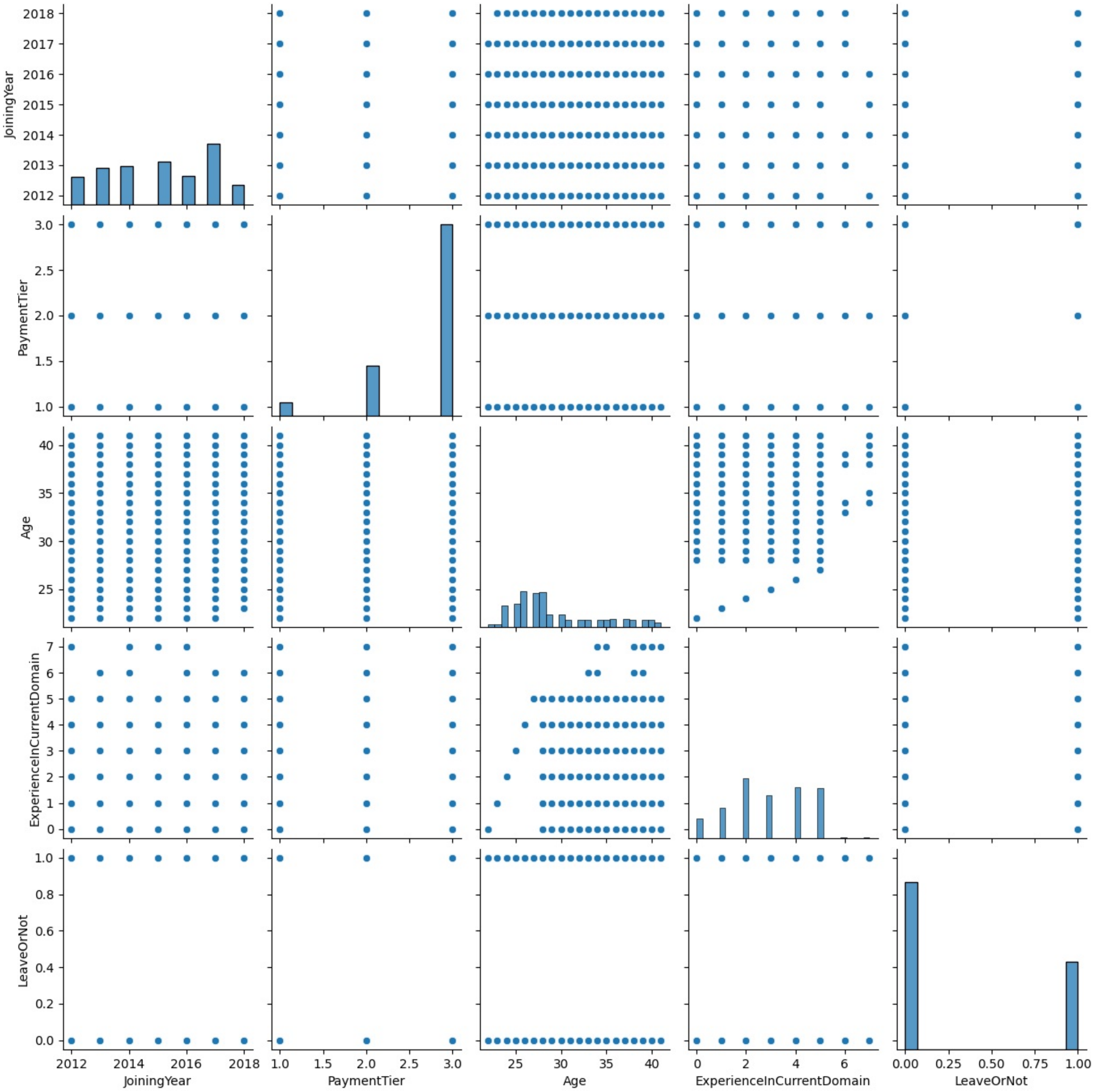
```
In [13]: df.value_counts()
```

```
Out[13]: Emp_id  Education  JoiningYear  City       PaymentTier  Age  Gender  EverBenched  ExperienceInCurrentDomain  Le
         aveOrNot
         E999    Bachelors  2015         Bangalore  3            28   Male    No           5                          0
         1
         E001    Bachelors  2017         Bangalore  3            34   Male    No           0                          0
         1
         E002    Bachelors  2013         Pune       1            28   Female  No           3                          1
         1
         E003    Bachelors  2014         New Delhi  3            38   Female  No           2                          0
         1
         E004    Masters    2016         Bangalore  3            27   Male    No           5                          1
         1

         ..
         E018    Bachelors  2014         Pune       3            34   Male    No           4                          0
         1
         E017    Bachelors  2014         Bangalore  3            34   Female  No           2                          0
         1
         E016    Bachelors  2017         Bangalore  1            29   Male    No           3                          0
         1
         E015    Bachelors  2012         Bangalore  3            37   Male    No           4                          0
         1
         E014    Bachelors  2016         Bangalore  3            39   Male    No           2                          0
         1
         Name: count, Length: 4653, dtype: int64
```

Import Seaborn

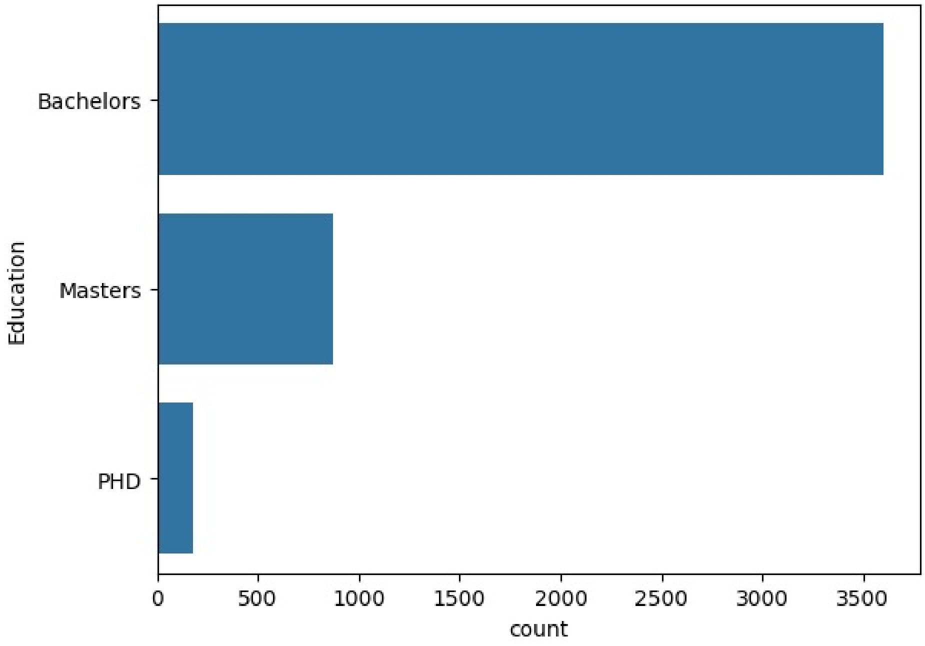
```
In [20]: import seaborn as sns
         sns.pairplot(df)
```

```
Out[20]: <seaborn.axisgrid.PairGrid at 0x15ecba33a90>
```



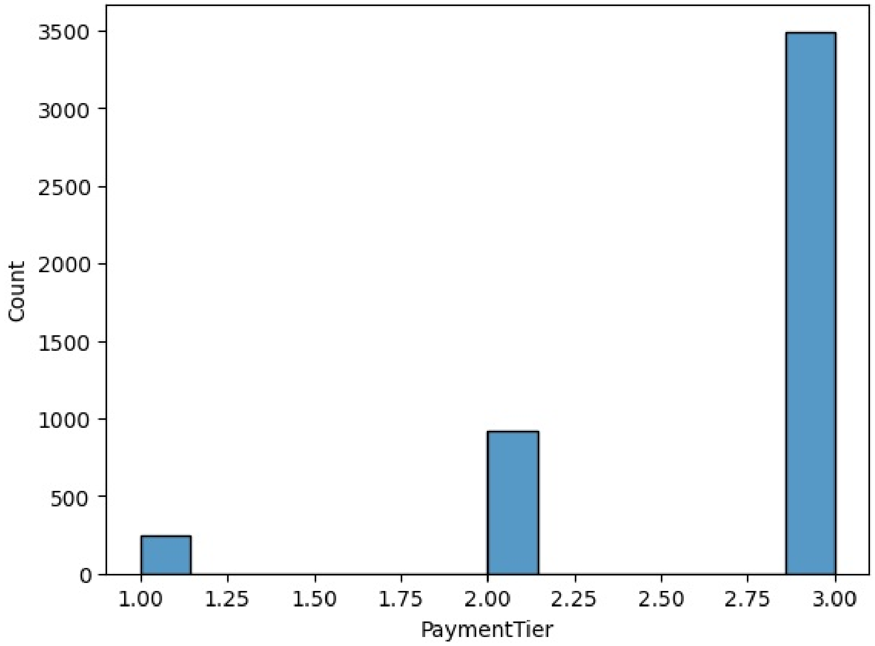
```
In [25]: sns.countplot(df["Education"])
```

```
Out[25]: <Axes: xlabel='count', ylabel='Education'>
```



In [37]:
```
sns.histplot(df["PaymentTier"])
```

Out[37]:
```
<Axes: xlabel='PaymentTier', ylabel='Count'>
```

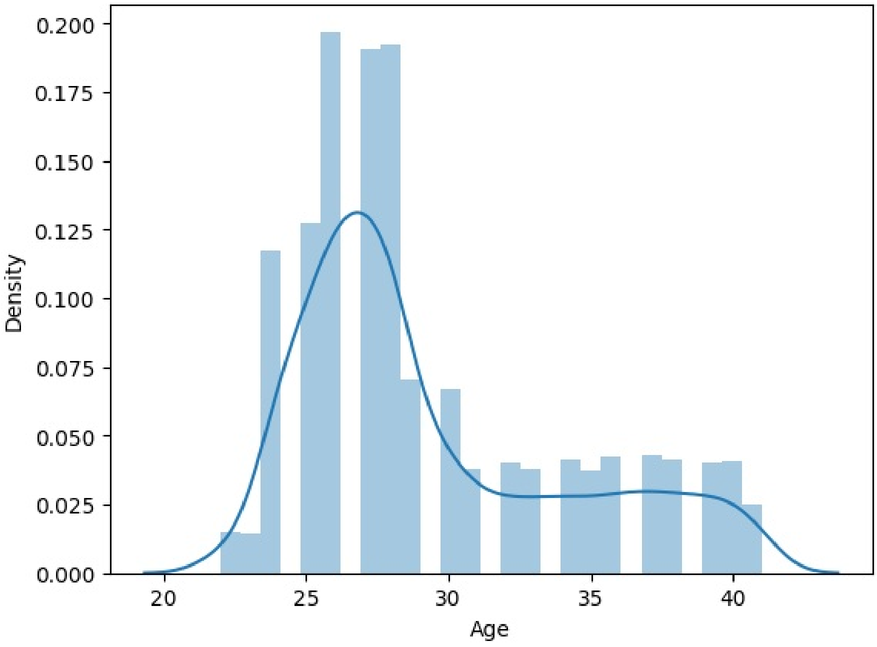


In [39]:
```
sns.distplot(df["Age"])
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_2936\2732350774.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(df["Age"])
```

Out[39]:

```
<Axes: xlabel='Age', ylabel='Density'>
```



Import Matplotlib

In [30]:

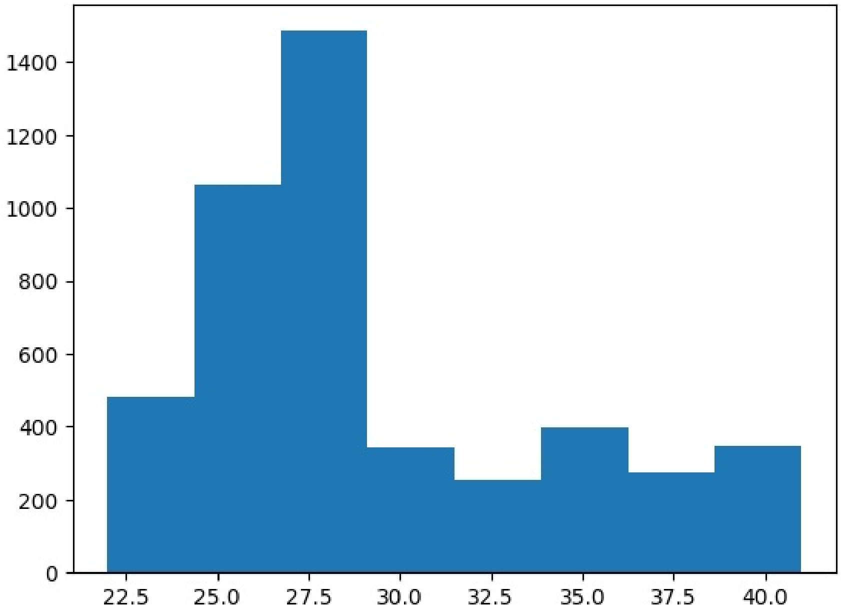
```
import matplotlib.pyplot as plt
```

In [35]:

```
plt.hist(df["Age"],bins=8)
```

Out[35]:

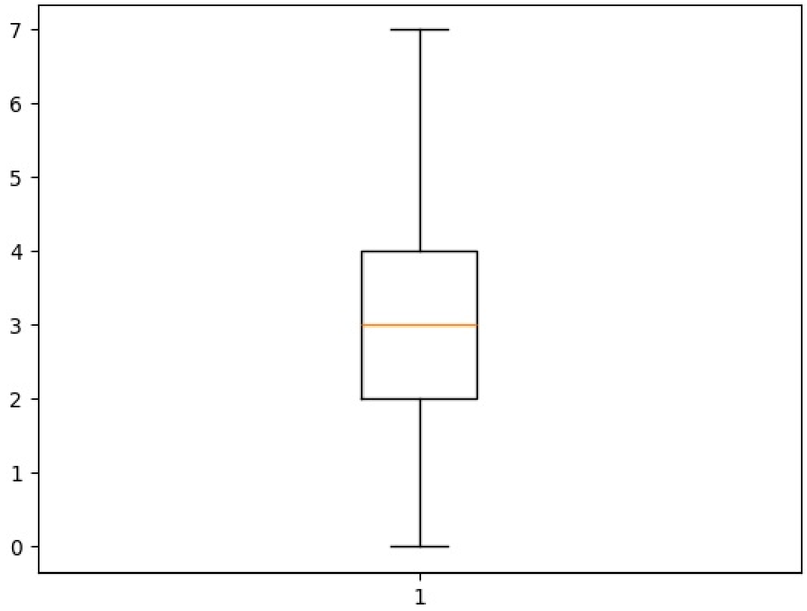
```
(array([ 482., 1063., 1485.,  345.,  256.,  398.,  277.,  347.]),
 array([22.   , 24.375, 26.75 , 29.125, 31.5  , 33.875, 36.25 , 38.625,
        41.   ]),
 <BarContainer object of 8 artists>)
```



In [41]:

```
plt.boxplot(df["ExperienceInCurrentDomain"])
```

```
Out[41]: {'whiskers': [<matplotlib.lines.Line2D at 0x15ed6d01690>,
           <matplotlib.lines.Line2D at 0x15ed6d01990>],
          'caps': [<matplotlib.lines.Line2D at 0x15ed6d01c90>,
           <matplotlib.lines.Line2D at 0x15ed6d01e70>],
          'boxes': [<matplotlib.lines.Line2D at 0x15ed6d01390>],
          'medians': [<matplotlib.lines.Line2D at 0x15ed6d02170>],
          'fliers': [<matplotlib.lines.Line2D at 0x15ed6d02470>],
          'means': []}
```



Summary of the EDA Report

1. **Data Overview**
   - `.info()` and `.describe()` provided details of data types, counts, and statistical summaries.
   - The dataset contains primarily **numerical columns** with some categorical fields, suitable for correlation and distribution analysis.
2. **Univariate Analysis**
   - **Age (Histogram):** Most individuals are in the **24–30 years range**, with frequency dropping at higher ages.
   - **Experience in Current Domain (Boxplot):** The majority have a moderate level of experience, but several **outliers** suggest highly experienced individuals.
3. **Bivariate/Multivariate Analysis**
   - **Pairplot:** Showed pairwise relationships between numerical variables, highlighting trends and possible group separations when colored by categories.
   - **Heatmap (Correlation Matrix):** Revealed the strength of relationships among numerical variables, helping to identify which features are strongly or weakly correlated.
   - **Scatterplots:** Used to visualize relationships between pairs of variables, showing patterns and potential clusters.
4. **Key Insights**
   - The dataset is dominated by **young professionals** with relatively lower years of domain experience.
   - Outliers in experience indicate the presence of a few senior-level individuals.
   - Correlation analysis shows which variables move together, which can be useful for predictive modeling or deeper analysis.

Overall Conclusion

The EDA provided a clear picture of the dataset’s structure, distribution, and relationships. It highlights that while most of the population is early in their careers, there is diversity in experience levels. These insights set a strong foundation for further analysis or modeling.