In [1]:
```python
# Importing Libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```python
# loading and reading dataset

employee = pd.read_csv("Desktop/Datasets/case-study-hr-analytics-in-power-bi/Dat
employee
```

Out[2]:

	ï»¿EmployeeID	FirstName	LastName	Gender	Age	BusinessTravel	Departm
0	3012-1A41	Leonelle	Simco	Female	30	Some Travel	S
1	CBCB-9C9D	Leonerd	Aland	Male	38	Some Travel	S
2	95D7-1CE9	Ahmed	Sykes	Male	43	Some Travel	Hur Resou
3	47A0-559B	Ermentrude	Berrie	Non-Binary	39	Some Travel	Technol
4	42CC-040A	Stace	Savege	Female	29	Some Travel	Hur Resou
...	...	...	...	...	...	...	
1465	467E-977A	Jud	Melanaphy	Male	20	Some Travel	Technol
1466	6FB9-A624	Marc	Calver	Non-Binary	27	Some Travel	Technol
1467	EBF4-5928	Rudolph	MacDearmont	Male	21	Some Travel	S
1468	60E6-B1D9	Merill	Agg	Male	21	Some Travel	Technol
1469	84D4-D4C3	Naoma	Hebbard	Female	20	No Travel	Technol

1470 rows × 23 columns

In [3]:
```
# Change column'name: 'ï»¿EmployeeID' to 'EmployeeID'

employee.rename(columns = {'ï»¿EmployeeID':'EmployeeID'}, inplace = True)
```

In [4]:
```
# Check if column name changed

employee.columns
```

Out[4]:
```
Index(['EmployeeID', 'FirstName', 'LastName', 'Gender', 'Age',
       'BusinessTravel', 'Department', 'DistanceFromHome (KM)', 'State',
       'Ethnicity', 'Education', 'EducationField', 'JobRole', 'MaritalStatus',
       'Salary', 'StockOptionLevel', 'OverTime', 'HireDate', 'Attrition',
       'YearsAtCompany', 'YearsInMostRecentRole', 'YearsSinceLastPromotion',
       'YearsWithCurrManager'],
      dtype='object')
```

In [5]:
```
# Overall information about this dataset

employee.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 23 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   EmployeeID                1470 non-null   object
 1   FirstName                 1470 non-null   object
 2   LastName                  1470 non-null   object
 3   Gender                    1470 non-null   object
 4   Age                       1470 non-null   int64
 5   BusinessTravel            1470 non-null   object
 6   Department                1470 non-null   object
 7   DistanceFromHome (KM)     1470 non-null   int64
 8   State                     1470 non-null   object
 9   Ethnicity                 1470 non-null   object
 10  Education                 1470 non-null   int64
 11  EducationField            1470 non-null   object
 12  JobRole                   1470 non-null   object
 13  MaritalStatus             1470 non-null   object
 14  Salary                    1470 non-null   int64
 15  StockOptionLevel          1470 non-null   int64
 16  OverTime                  1470 non-null   object
 17  HireDate                  1470 non-null   object
 18  Attrition                 1470 non-null   object
 19  YearsAtCompany            1470 non-null   int64
 20  YearsInMostRecentRole     1470 non-null   int64
 21  YearsSinceLastPromotion   1470 non-null   int64
 22  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(9), object(14)
memory usage: 264.3+ KB
```

In [6]:
```
# Dataset statistics for numerical columns

employee.describe()
```

Out[6]:

	Age	DistanceFromHome (KM)	Education	Salary	StockOptionLevel
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	28.989796	22.502721	2.912925	112956.497959	0.793878
std	7.993055	12.811124	1.024165	103342.889222	0.852077
min	18.000000	1.000000	1.000000	20387.000000	0.000000
25%	23.000000	12.000000	2.000000	43580.500000	0.000000
50%	26.000000	22.000000	3.000000	71199.500000	1.000000
75%	34.000000	33.000000	4.000000	142055.750000	1.000000
max	51.000000	45.000000	5.000000	547204.000000	3.000000

In [7]:
```
# Check if there is missing values

employee.isnull().sum()
```

Out[7]:
```
EmployeeID                0
FirstName                 0
LastName                  0
Gender                    0
Age                       0
BusinessTravel            0
Department                0
DistanceFromHome (KM)     0
State                     0
Ethnicity                 0
Education                 0
EducationField            0
JobRole                   0
MaritalStatus             0
Salary                    0
StockOptionLevel          0
OverTime                  0
HireDate                  0
Attrition                 0
YearsAtCompany            0
YearsInMostRecentRole     0
YearsSinceLastPromotion   0
YearsWithCurrManager      0
dtype: int64
```

In [8]:
```
# Check duplicates

employee.duplicated().sum()
```

Out[8]:
```
0
```

In [9]:
```
# Check top rows of dataset

employee.head()
```

Out[9]:

	EmployeeID	FirstName	LastName	Gender	Age	BusinessTravel	Department	Dist
0	3012-1A41	Leonelle	Simco	Female	30	Some Travel	Sales	
1	CBCB-9C9D	Leonerd	Aland	Male	38	Some Travel	Sales	
2	95D7-1CE9	Ahmed	Sykes	Male	43	Some Travel	Human Resources	
3	47A0-559B	Ermentrude	Berrie	Non-Binary	39	Some Travel	Technology	
4	42CC-040A	Stace	Savege	Female	29	Some Travel	Human Resources	

5 rows × 23 columns

In [10]:

```
# Check bottom rows of dataset

employee.tail()
```

Out[10]:

	EmployeeID	FirstName	LastName	Gender	Age	BusinessTravel	Department
1465	467E-977A	Jud	Melanaphy	Male	20	Some Travel	Technology
1466	6FB9-A624	Marc	Calver	Non-Binary	27	Some Travel	Technology
1467	EBF4-5928	Rudolph	MacDearmont	Male	21	Some Travel	Sales
1468	60E6-B1D9	Merill	Agg	Male	21	Some Travel	Technology
1469	84D4-D4C3	Naoma	Hebbard	Female	20	No Travel	Technology

5 rows × 23 columns

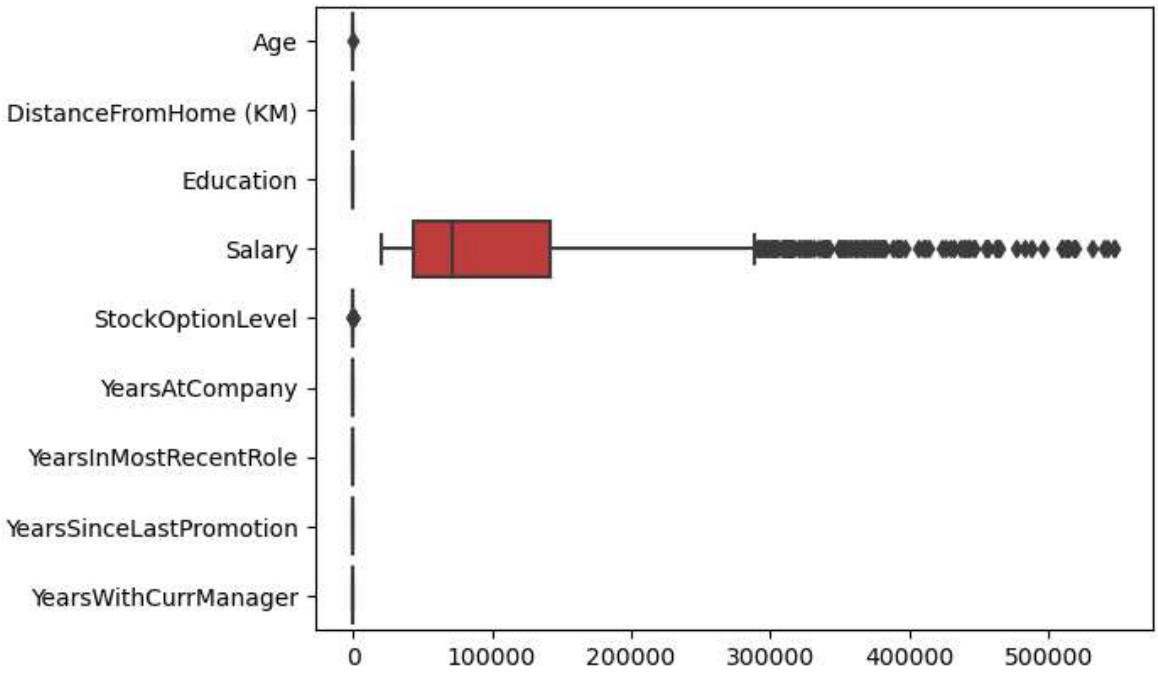
In [11]:

```
#Check outliers using Boxplot for numerical columns

sns.boxplot(employee,orient='h')
```

Out[11]:

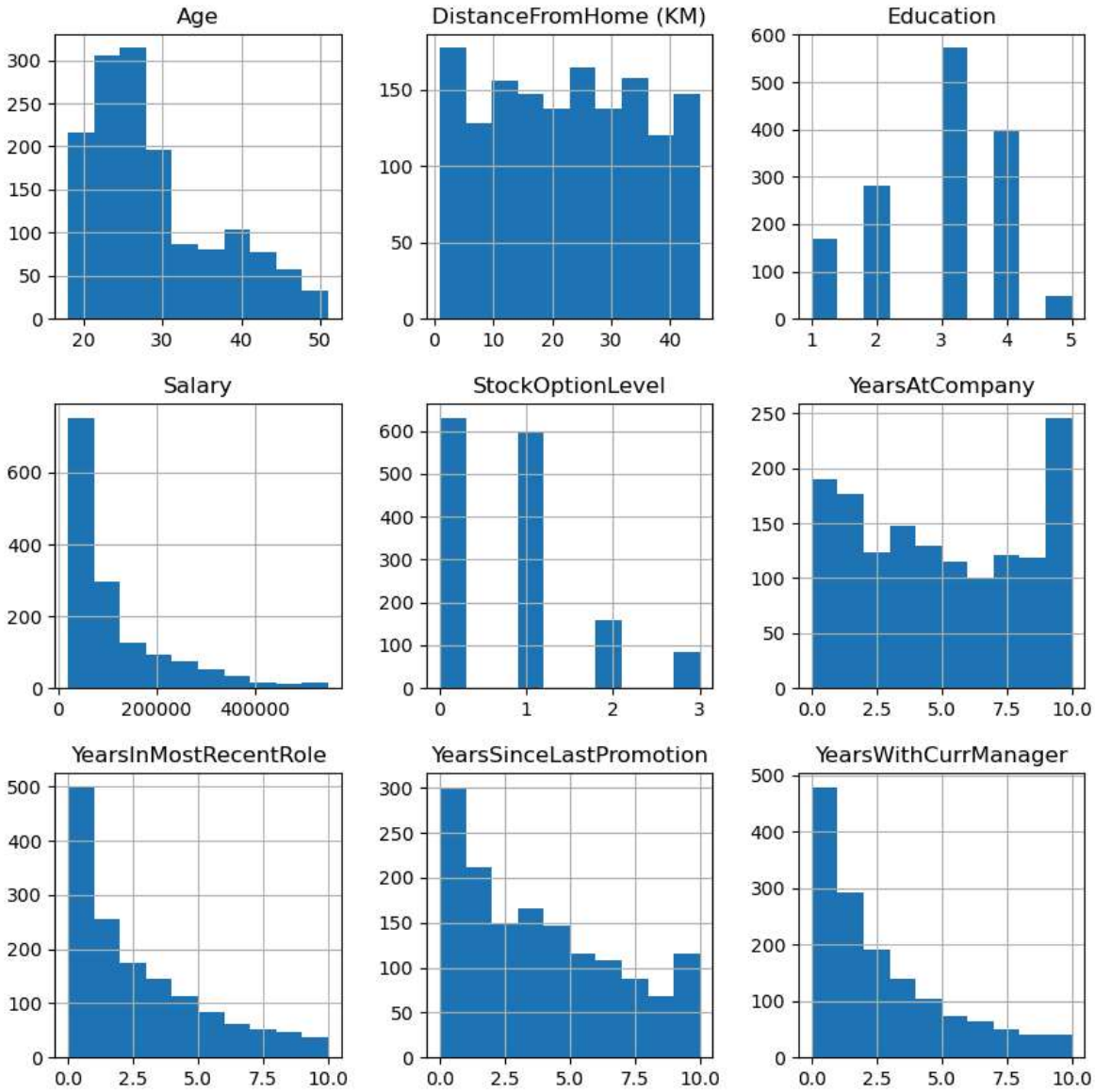
```
<Axes: >
```



In [12]:
```
# check data distribution using histograms for numerical columns

employee.hist(figsize=(10,10))
```

Out[12]:
```
array([[<Axes: title={'center': 'Age'}>,
        <Axes: title={'center': 'DistanceFromHome (KM)'}>,
        <Axes: title={'center': 'Education'}>],
       [<Axes: title={'center': 'Salary'}>,
        <Axes: title={'center': 'StockOptionLevel'}>,
        <Axes: title={'center': 'YearsAtCompany'}>],
       [<Axes: title={'center': 'YearsInMostRecentRole'}>,
        <Axes: title={'center': 'YearsSinceLastPromotion'}>,
        <Axes: title={'center': 'YearsWithCurrManager'}>]], dtype=object)
```

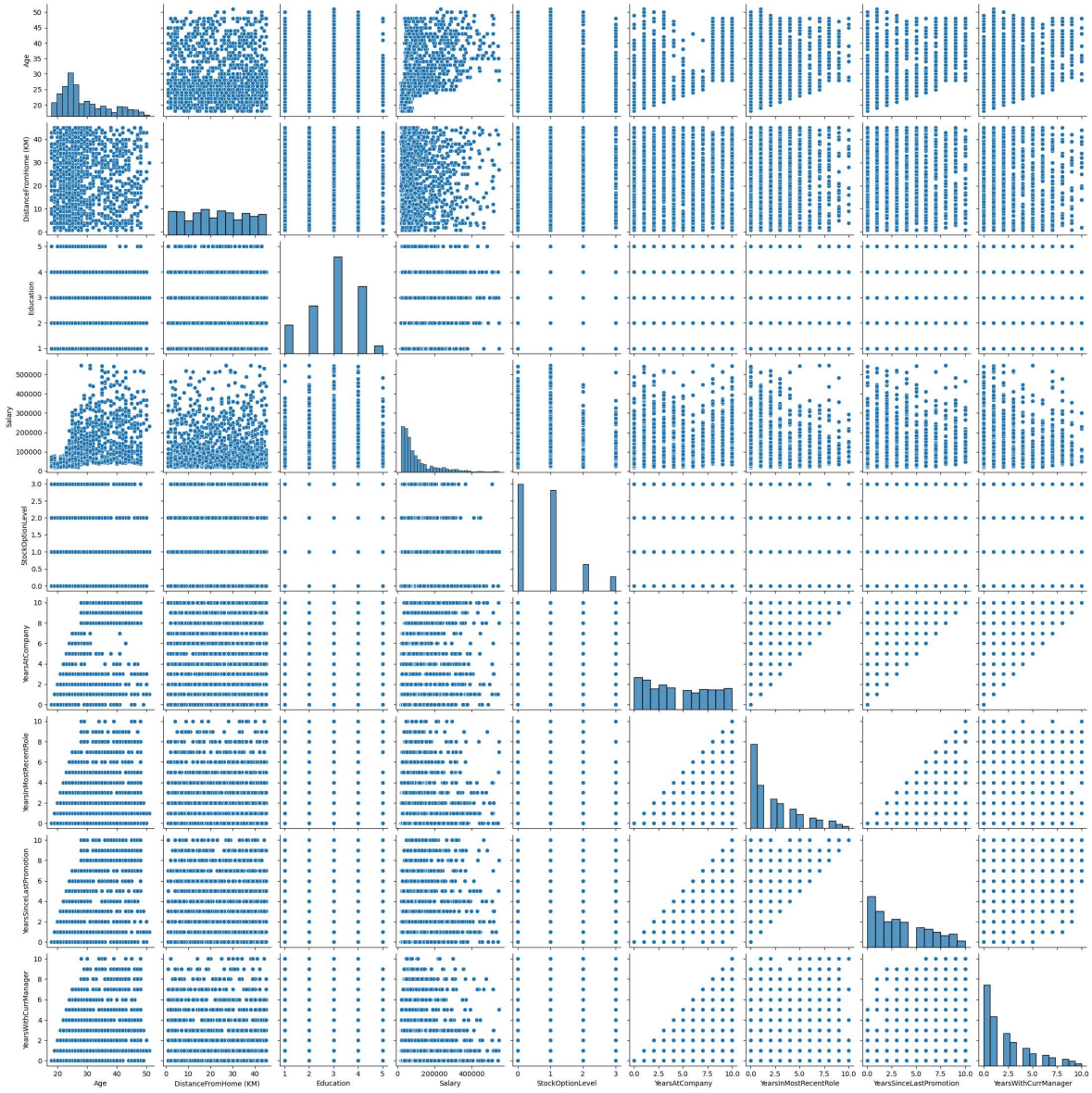


In [13]:
```
# Explore the relationships between variables

sns.pairplot(employee)
```

```
C:\Users\thaop\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning:
The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

Out[13]:
```
<seaborn.axisgrid.PairGrid at 0x20206852090>
```



In [14]:
```
employee.to_csv('Atlas Employee.csv')
```

In [ ]: