In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
data=pd.read_csv('HR_ANALYTICS.csv')
```

In [3]:
```
data.head()
```

Out[3]:

| | Candidate Ref | DOJ Extended | D-DOJ Extended | Duration to accept offer | Notice period | Offered band | E0 | E1 | E2 | E3 | ... | D-AXON | Location |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2110407 | Yes | 1 | 14.0 | 30 | E2 | 0 | 0 | 1 | 0 | ... | 0 | Noida |
| **1** | 2112635 | No | 0 | 18.0 | 30 | E2 | 0 | 0 | 1 | 0 | ... | 0 | Chennai |
| **2** | 2112838 | No | 0 | 3.0 | 45 | E2 | 0 | 0 | 1 | 0 | ... | 0 | Noida |
| **3** | 2115021 | No | 0 | 26.0 | 30 | E2 | 0 | 0 | 1 | 0 | ... | 0 | Noida |
| **4** | 2115125 | Yes | 1 | 1.0 | 120 | E2 | 0 | 0 | 1 | 0 | ... | 0 | Noida |

5 rows × 37 columns

In [4]:
```
data.isnull().sum().sum()
```

Out[4]:
```
5343
```

In [5]:
```
data.shape
```

Out[5]:
```
(12333, 37)
```

In [6]:
```
data.describe()
```

Out[6]:

| | Candidate Ref | D-DOJ Extended | Duration to accept offer | Notice period | E0 | E1 | |
|---|---|---|---|---|---|---|---|
| **count** | 1.233300e+04 | 12333.000000 | 9614.000000 | 12333.000000 | 12333.000000 | 12333.000000 | 12333.0000 |
| **mean** | 2.872888e+06 | 0.411417 | 21.189619 | 37.943323 | 0.085786 | 0.567259 | 0.2805 |
| **std** | 5.099726e+05 | 0.492110 | 26.410351 | 24.526587 | 0.280059 | 0.495476 | 0.4492 |
| **min** | 2.109586e+06 | 0.000000 | -228.000000 | 0.000000 | 0.000000 | 0.000000 | 0.0000 |
| **25%** | 2.378256e+06 | 0.000000 | 2.000000 | 30.000000 | 0.000000 | 0.000000 | 0.0000 |
| **50%** | 2.820675e+06 | 0.000000 | 9.000000 | 30.000000 | 0.000000 | 1.000000 | 0.0000 |
| **75%** | 3.338197e+06 | 1.000000 | 32.000000 | 60.000000 | 0.000000 | 1.000000 | 1.0000 |
| **max** | 3.836076e+06 | 1.000000 | 224.000000 | 120.000000 | 1.000000 | 1.000000 | 1.0000 |

8 rows × 28 columns

In [7]:
```
data.columns
```

Out[7]:
```
Index(['Candidate Ref', 'DOJ Extended', 'D-DOJ Extended ',
       'Duration to accept offer', 'Notice period', 'Offered band', 'E0', 'E1',
       'E2', 'E3', 'E4', 'E5', 'Pecent hike expected in CTC',
       'Percent hike offered in CTC', 'Percent difference CTC',
       'Joining Bonus', 'D-Joining Bonus', 'Candidate relocate actual',
       'Gender', 'Candidate Source', 'D-Agency', 'D-Employee Referral',
       'Rex in Yrs', 'LOB', 'D-Sales', 'D-Healthcare', 'D-Infra', 'D-AXON',
       'Location', 'D-Bangalore', 'D-Chennai', 'D-Mumbai', 'D-Delhi NCR',
       'D-Hyderabad', 'Age', 'Status', 'D-Joining'],
      dtype='object')
```

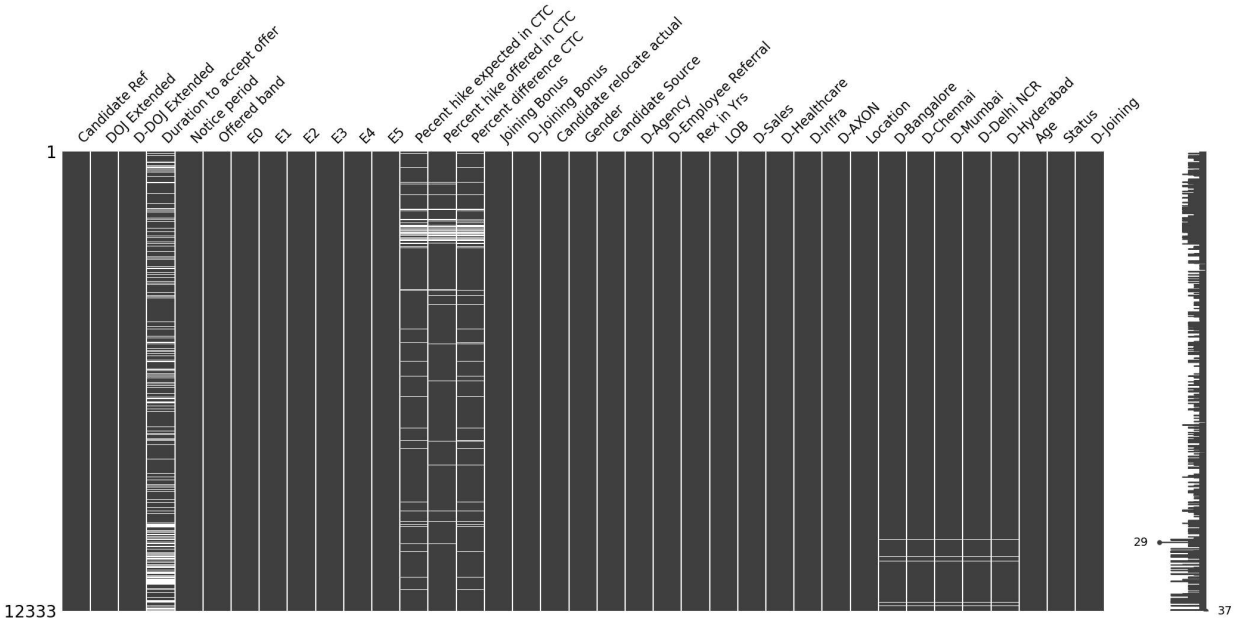
In [8]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12333 entries, 0 to 12332
Data columns (total 37 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   Candidate Ref                 12333 non-null  int64  
 1   DOJ Extended                  12333 non-null  object 
 2   D-DOJ Extended                12333 non-null  int64  
 3   Duration to accept offer      9614 non-null   float64
 4   Notice period                 12333 non-null  int64  
 5   Offered band                  12333 non-null  object 
 6   E0                            12333 non-null  int64  
 7   E1                            12333 non-null  int64  
 8   E2                            12333 non-null  int64  
 9   E3                            12333 non-null  int64  
 10  E4                            12333 non-null  int64  
 11  E5                            12333 non-null  int64  
 12  Pecent hike expected in CTC   11586 non-null  float64
 13  Percent hike offered in CTC   11737 non-null  float64
 14  Percent difference CTC        11482 non-null  float64
 15  Joining Bonus                 12333 non-null  object 
 16  D-Joining Bonus               12333 non-null  int64  
 17  Candidate relocate actual     12333 non-null  object 
 18  Gender                        12333 non-null  object 
 19  Candidate Source              12333 non-null  object 
 20  D-Agency                      12333 non-null  int64  
 21  D-Employee Referral           12333 non-null  int64  
 22  Rex in Yrs                    12333 non-null  int64  
 23  LOB                           12333 non-null  object 
 24  D-Sales                       12333 non-null  int64  
 25  D-Healthcare                  12333 non-null  int64  
 26  D-Infra                       12333 non-null  int64  
 27  D-AXON                        12333 non-null  int64  
 28  Location                      12333 non-null  object 
 29  D-Bangalore                   12247 non-null  float64
 30  D-Chennai                     12247 non-null  float64
 31  D-Mumbai                      12247 non-null  float64
 32  D-Delhi NCR                   12247 non-null  float64
 33  D-Hyderabad                   12247 non-null  float64
 34  Age                           12333 non-null  int64  
 35  Status                        12333 non-null  object 
 36  D-Joining                     12333 non-null  int64  
dtypes: float64(9), int64(19), object(9)
memory usage: 3.5+ MB
```

In [9]:
```
import missingno as msno
```

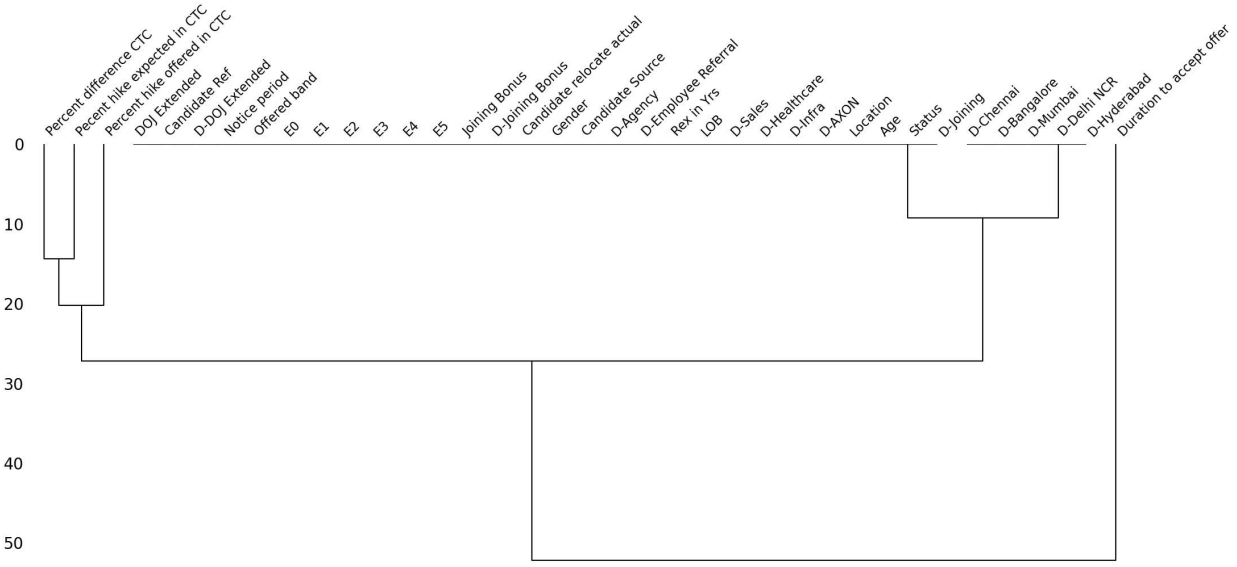
In [10]:
```
msno.matrix(data)
```

Out[10]:
```
<AxesSubplot:>
```



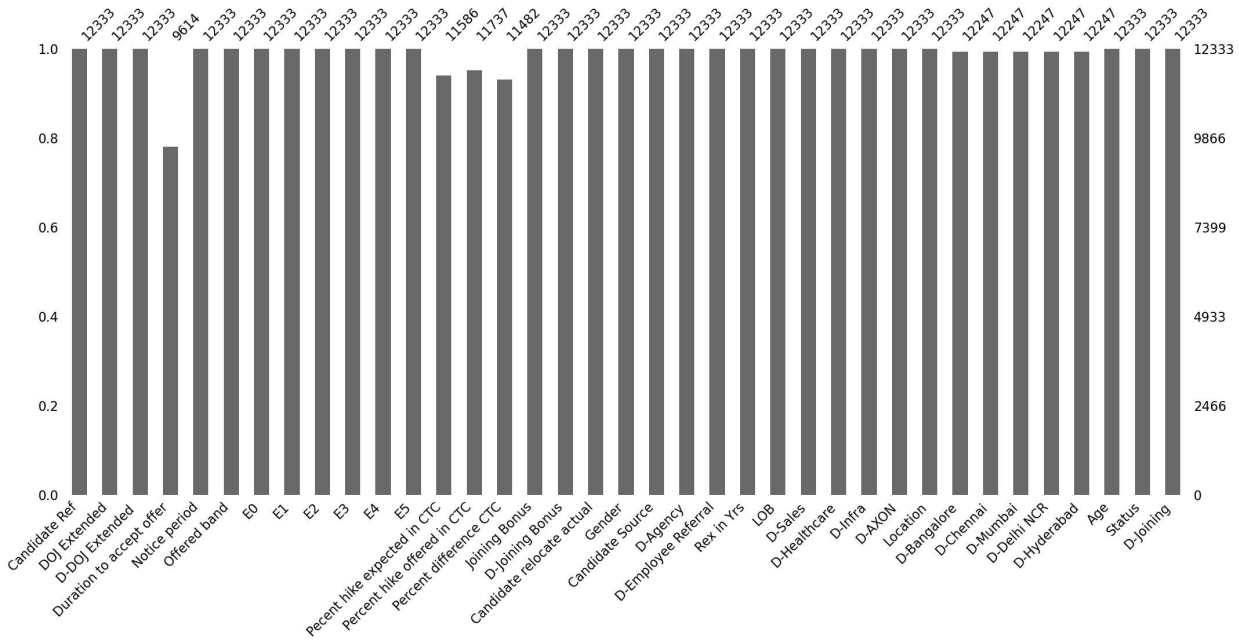

In [11]:
```
msno.dendrogram(data)
```

Out[11]: <AxesSubplot:>




In [12]:
```
msno.bar(data)
```

Out[12]: <AxesSubplot:>




In [15]:
```
data.dropna(how='all',inplace=True)
```

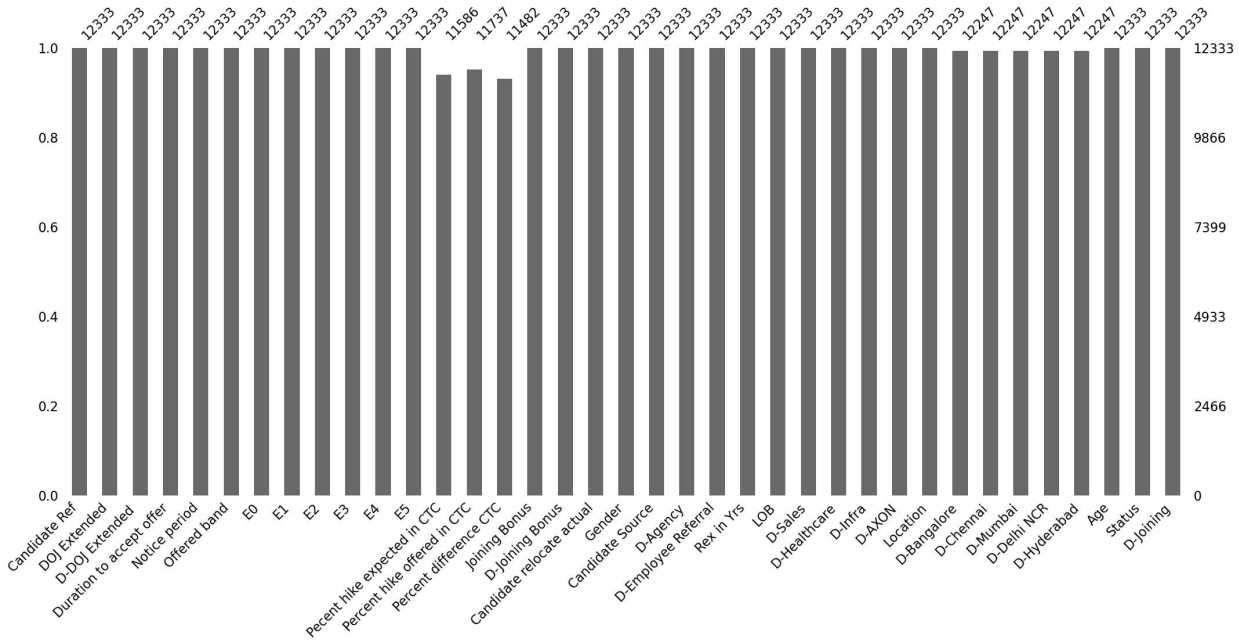
In [16]:
```
data['Duration to accept offer'].fillna(value=data['Duration to accept offer'].mean(),
```

In [17]:
```
data['Duration to accept offer'].isnull().sum().sum()
```

Out[17]:
```
0
```

In [18]:
```
msno.bar(data)
```

Out[18]:
```
<AxesSubplot:>
```




In [19]:
```
data.isnull().sum().sum()
```

Out[19]:
```
2624
```

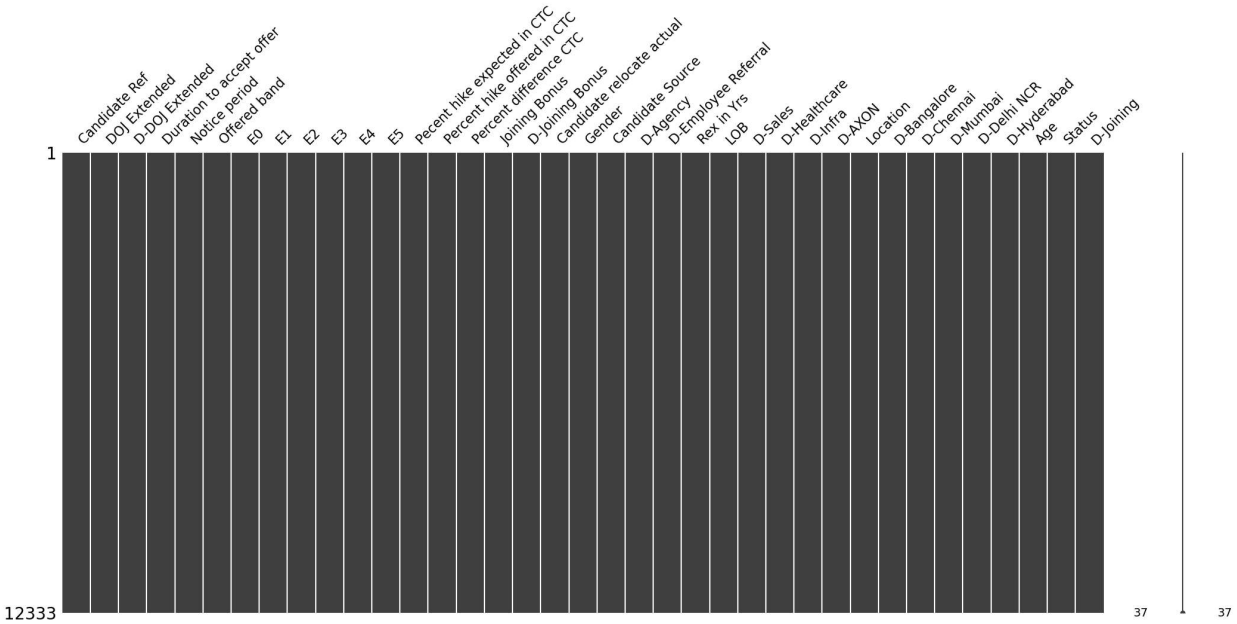
In [20]:
```
2624/12333*100
```

Out[20]:
```
21.276250709478635
```

In [21]:
```
data.fillna(value=data['Percent difference CTC'].median(),inplace=True)
```

In [22]:
```
msno.matrix(data)
```

Out[22]:
```
<AxesSubplot:>
```




In [23]:
```
data.dropna(inplace=True)
```

In [24]:
```
data.isnull().sum()
```

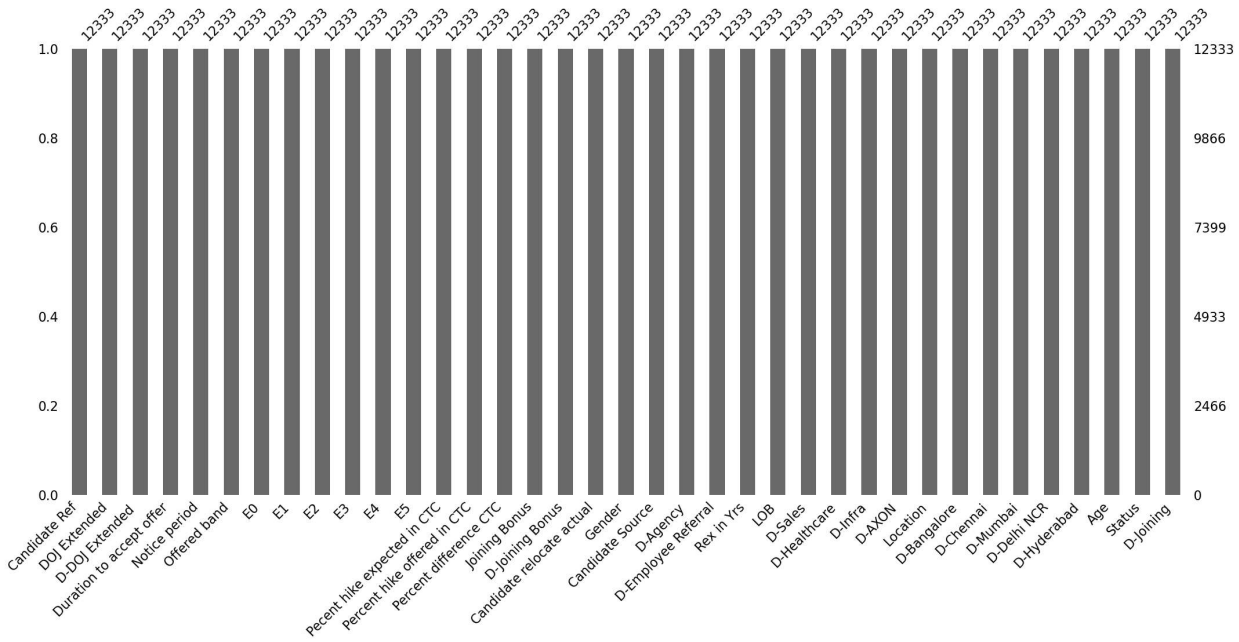
Out[24]:
```
Candidate Ref                0
DOJ Extended                 0
D-DOJ Extended               0
Duration to accept offer     0
Notice period                0
Offered band                 0
E0                           0
E1                           0
E2                           0
E3                           0
E4                           0
E5                           0
Pecent hike expected in CTC  0
Percent hike offered in CTC  0
Percent difference CTC       0
Joining Bonus                0
D-Joining Bonus              0
Candidate relocate actual    0
Gender                       0
Candidate Source             0
D-Agency                     0
D-Employee Referral          0
Rex in Yrs                   0
LOB                          0
D-Sales                      0
D-Healthcare                 0
D-Infra                      0
D-AXON                       0
Location                     0
D-Bangalore                  0
D-Chennai                    0
D-Mumbai                     0
D-Delhi NCR                  0
D-Hyderabad                  0
Age                          0
Status                       0
D-Joining                    0
dtype: int64
```

In [25]:
```
data.isnull().sum().sum()
```

Out[25]:
```
0
```

In [26]:
```
msno.bar(data)
```

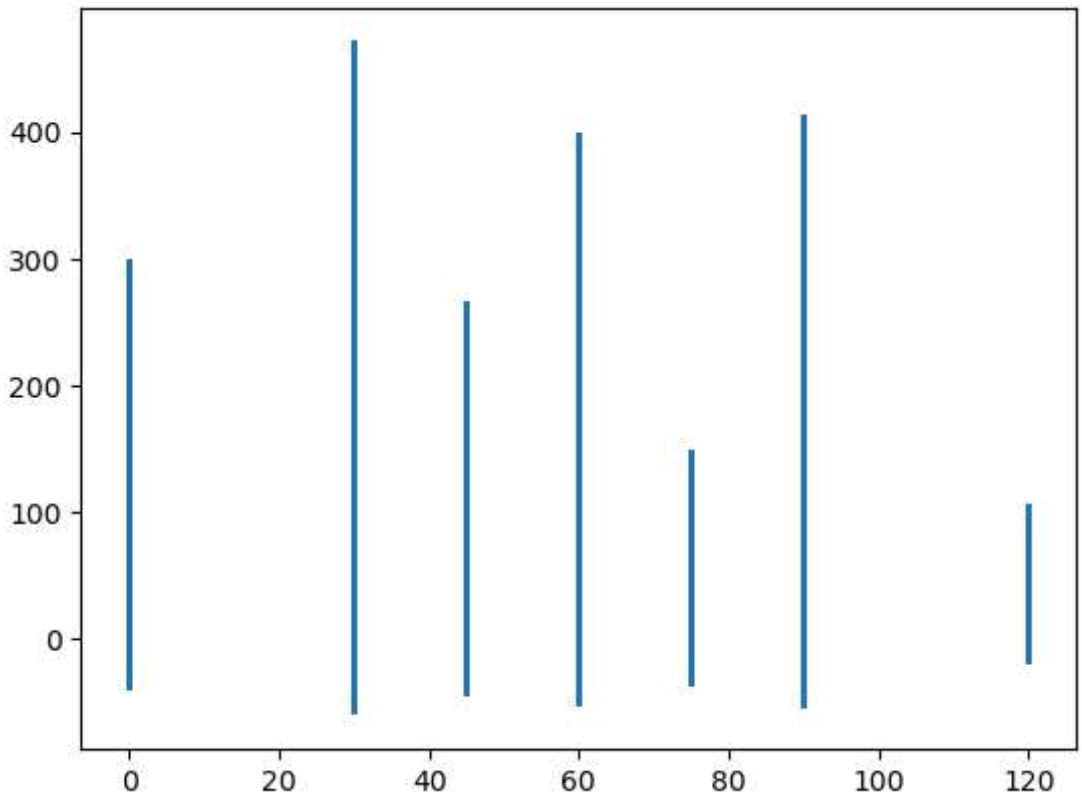
Out[26]:
```
<AxesSubplot:>
```




In [32]:
```
import matplotlib.pyplot as plt
%matplotlib inline
```

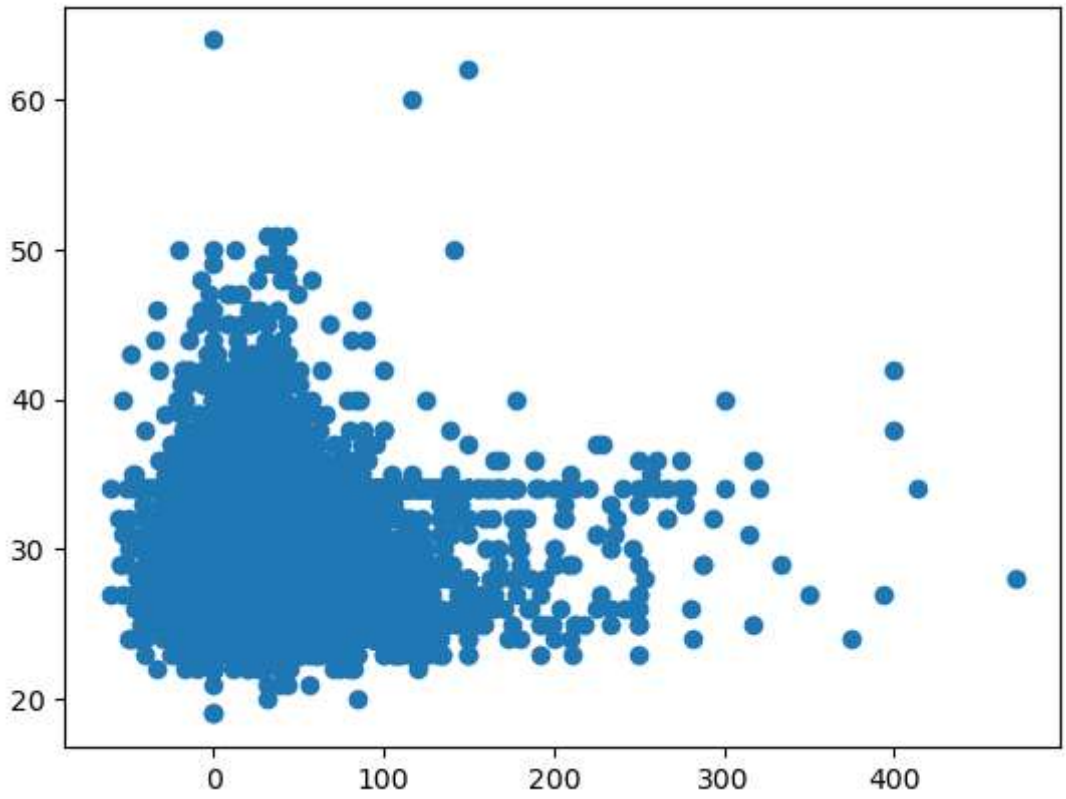
In [33]:
```
plt.bar(data['Notice period'],data['Percent hike offered in CTC'])
```

Out[33]:
```
<BarContainer object of 12333 artists>
```




In [34]:
```
plt.scatter(data['Percent hike offered in CTC'],data['Age'])
```

Out[34]:
```
<matplotlib.collections.PathCollection at 0x22cb40a6b80>
```

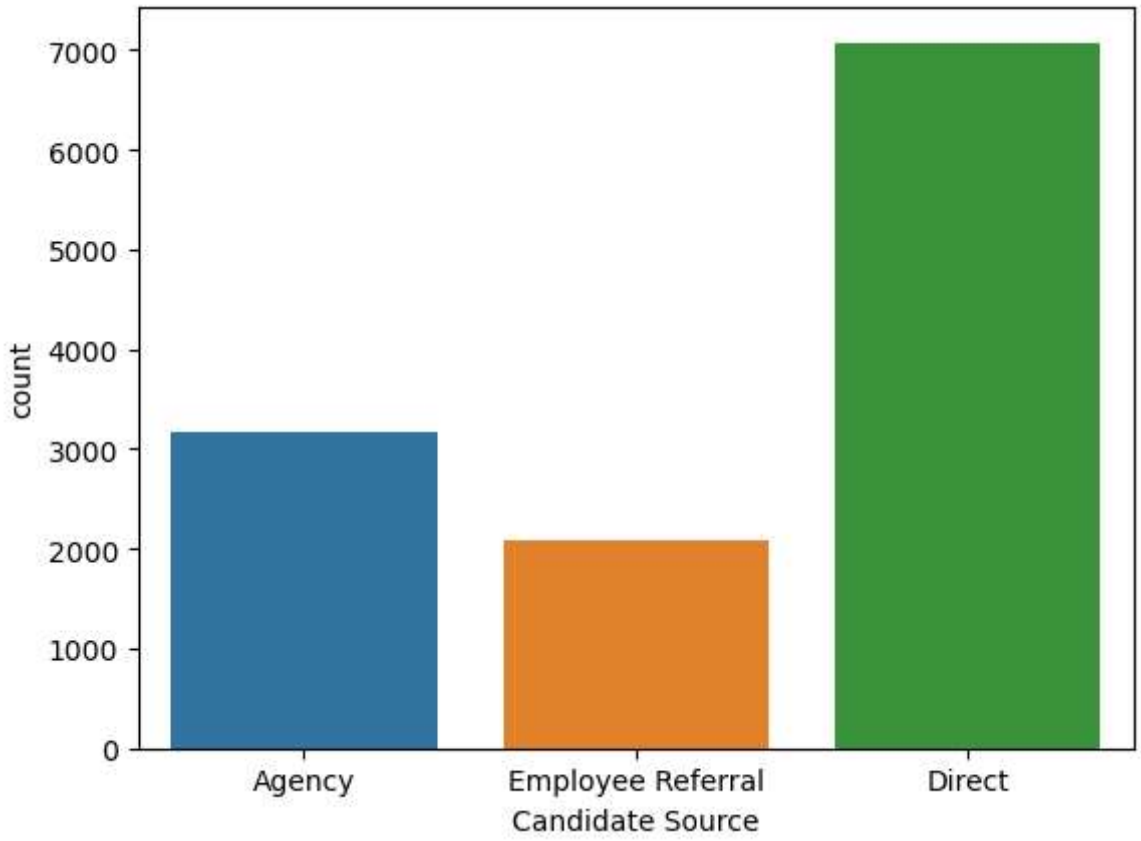



In [35]:
```
import seaborn as sns
```

In [36]:
```
sns.countplot(data['Candidate Source'])
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning:
Pass the following variable as a keyword arg: x. From version 0.12, the only valid po
sitional argument will be `data`, and passing other arguments without an explicit key
word will result in an error or misinterpretation.
  warnings.warn(
```

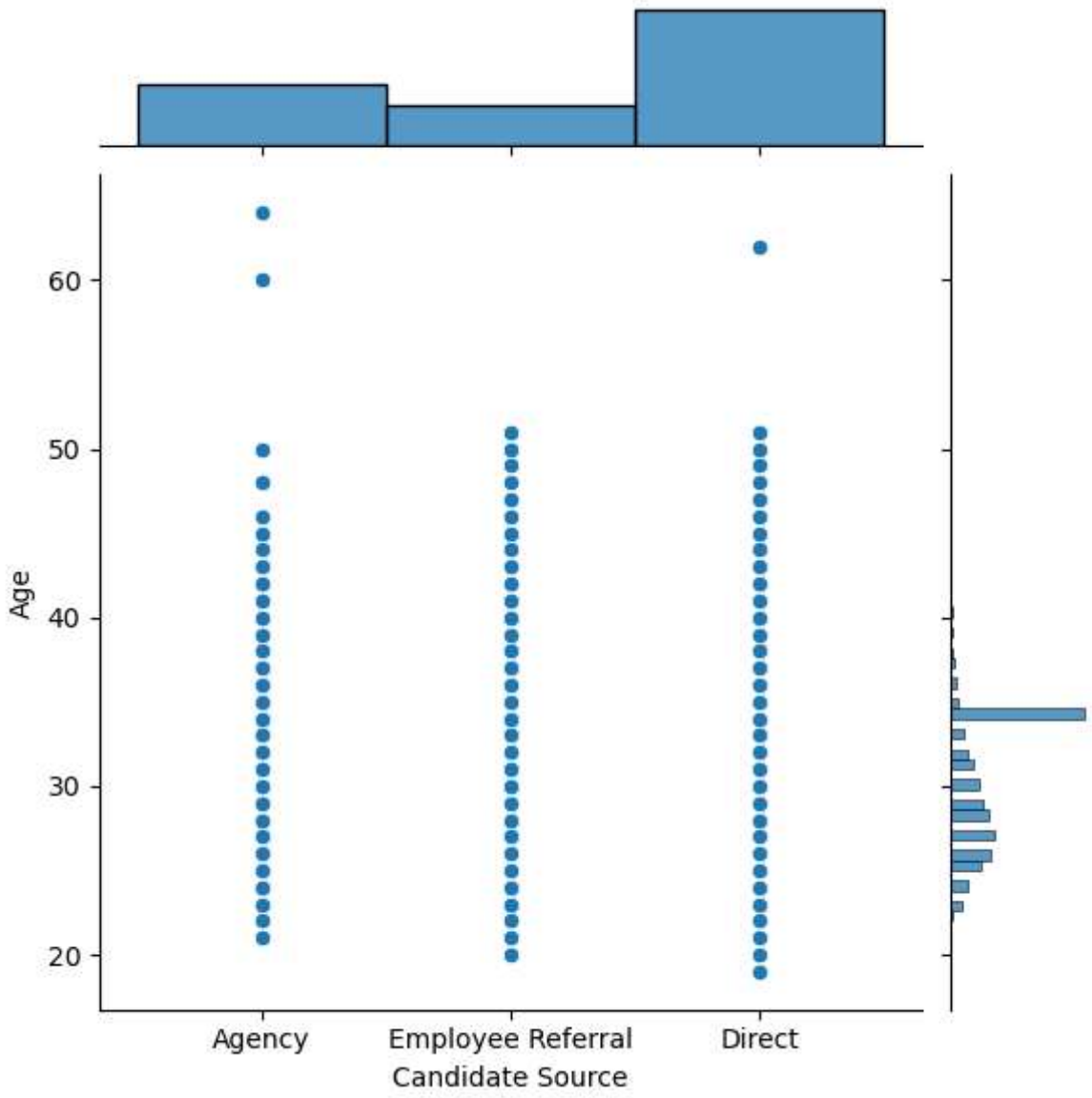
Out[36]:
```
<AxesSubplot:xlabel='Candidate Source', ylabel='count'>
```




In [37]:
```
sns.jointplot(data['Candidate Source'],data['Age'])
```

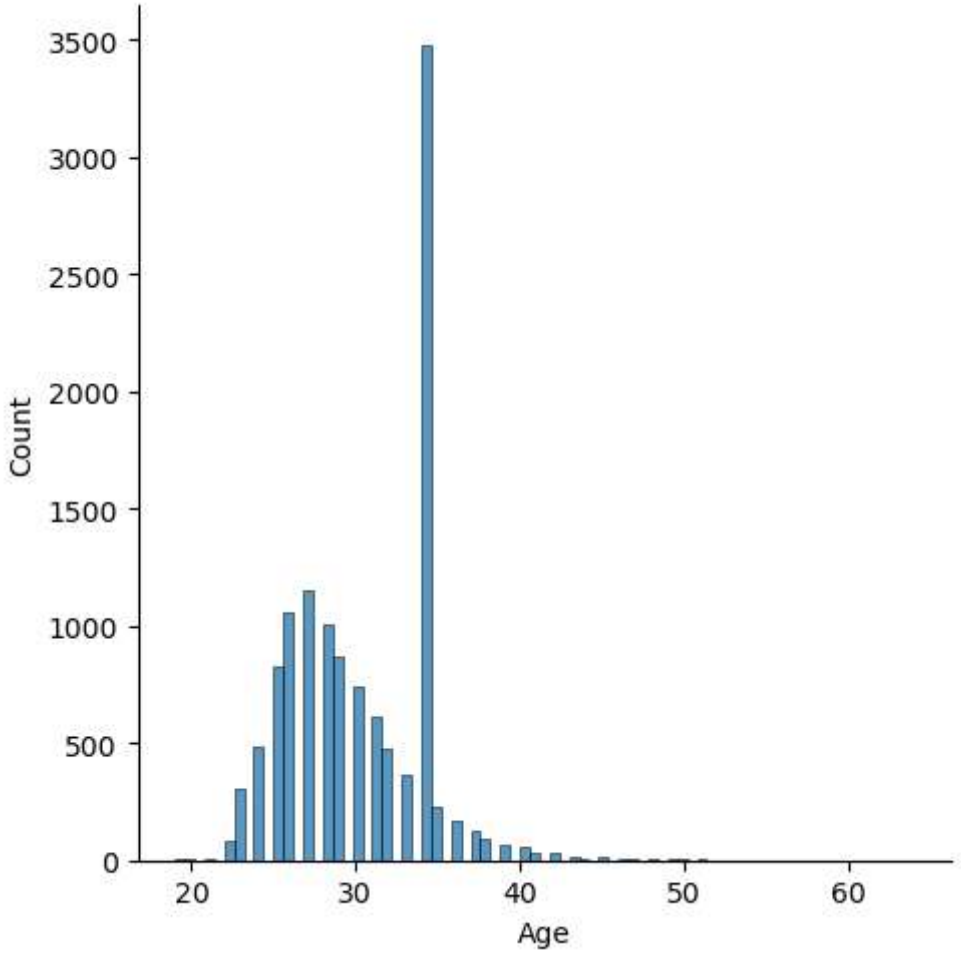
```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning:
Pass the following variables as keyword args: x, y. From version 0.12, the only valid
positional argument will be `data`, and passing other arguments without an explicit k
eyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[37]:
```
<seaborn.axisgrid.JointGrid at 0x22cb36be670>
```



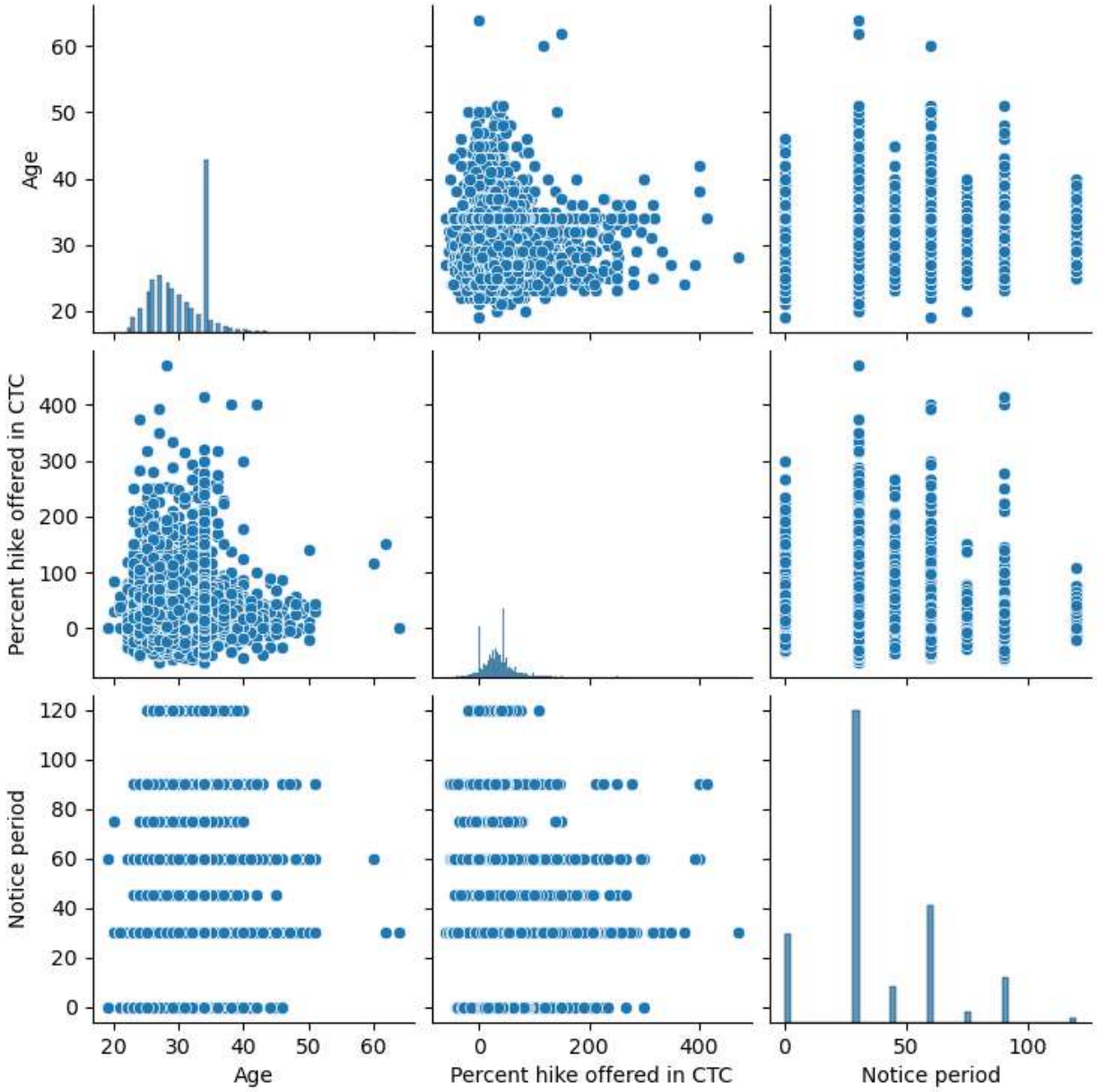

In [38]:
```
print(sns.displot(data['Age']))
```

```
<seaborn.axisgrid.FacetGrid object at 0x0000022CB36D75E0>
```




In [39]:
```
sns.pairplot(data[['Joining Bonus','Age','Percent hike offered in CTC','Notice period'
```

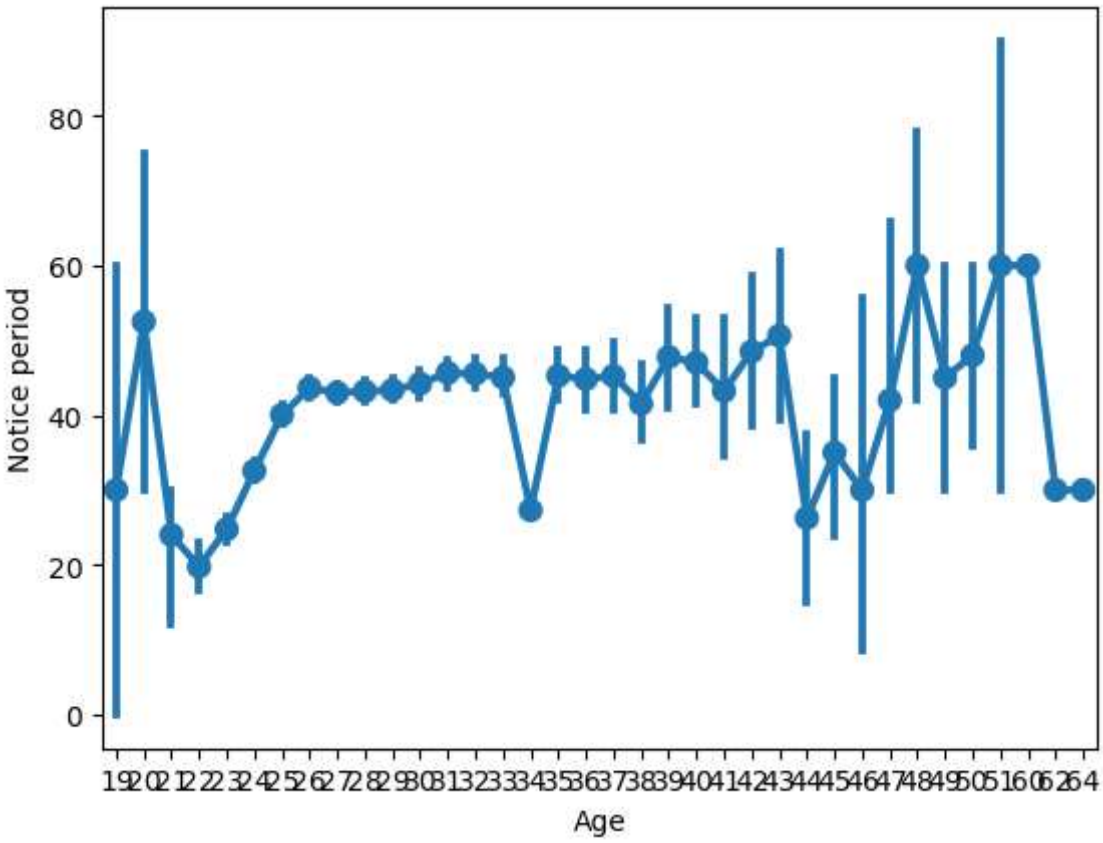
Out[39]:
```
<seaborn.axisgrid.PairGrid at 0x22cb3ee66a0>
```




In [40]:
```
sns.pointplot(data['Age'],data['Notice period'])
```

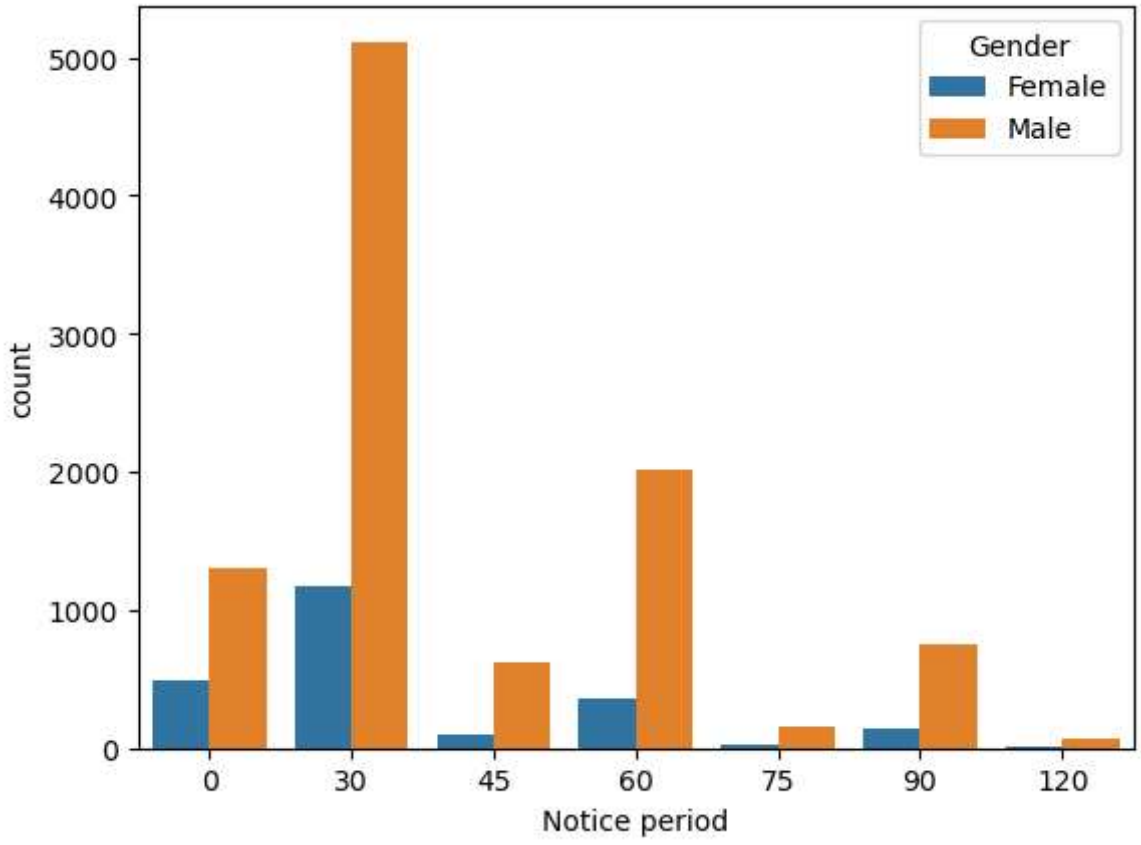
```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning:
Pass the following variables as keyword args: x, y. From version 0.12, the only valid
positional argument will be `data`, and passing other arguments without an explicit k
eyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[40]:
```
<AxesSubplot:xlabel='Age', ylabel='Notice period'>
```




In [50]:
```
# sns.countplot(data['GENDER'])
sns.countplot(x='Notice period',hue='Gender',data=data)
```

Out[50]:
```
<AxesSubplot:xlabel='Notice period', ylabel='count'>
```



In [55]:
```
data.corr()
```

Out[55]:

| | Candidate Ref | D-DOJ Extended | Duration to accept offer | Notice period | E0 | E1 | E2 | E3 |
|---|---|---|---|---|---|---|---|---|
| **Candidate Ref** | 1.000000 | 0.020973 | -2.258635e-02 | -0.221621 | 0.197311 | 0.039389 | -0.129199 | -0.066325 |
| **D-DOJ Extended** | 0.020973 | 1.000000 | 2.894843e-01 | 0.087780 | -0.079006 | 0.013878 | 0.030624 | 0.011559 |
| **Duration to accept offer** | -0.022586 | 0.289484 | 1.000000e+00 | 0.268292 | -0.021470 | -0.001566 | 0.010603 | 0.008861 |
| **Notice period** | -0.221621 | 0.087780 | 2.682923e-01 | 1.000000 | -0.391927 | 0.047667 | 0.130553 | 0.102843 |
| **E0** | 0.197311 | -0.079006 | -2.147015e-02 | -0.391927 | 1.000000 | -0.350720 | -0.191288 | -0.074572 |
| **E1** | 0.039389 | 0.013878 | -1.566345e-03 | 0.047667 | -0.350720 | 1.000000 | -0.714956 | -0.278720 |
| **E2** | -0.129199 | 0.030624 | 1.060260e-02 | 0.130553 | -0.191288 | -0.714956 | 1.000000 | -0.152018 |
| **E3** | -0.066325 | 0.011559 | 8.860922e-03 | 0.102843 | -0.074572 | -0.278720 | -0.152018 | 1.000000 |
| **E4** | -0.013528 | -0.017154 | -1.201591e-04 | 0.037535 | -0.028522 | -0.106603 | -0.058143 | -0.022667 |
| **E5** | -0.005457 | 0.009691 | -9.284021e-19 | 0.008989 | -0.013241 | -0.049489 | -0.026992 | -0.010523 |
| **Pecent hike expected in CTC** | 0.056408 | 0.034182 | 1.797153e-02 | 0.039889 | -0.141067 | 0.201650 | -0.066421 | -0.105078 |
| **Percent hike offered in CTC** | 0.043651 | 0.043864 | 1.441420e-02 | -0.008932 | -0.060857 | 0.049363 | 0.004754 | -0.018786 |
| **Percent difference CTC** | 0.004026 | 0.030180 | -6.851190e-03 | -0.056372 | 0.069255 | -0.141264 | 0.067717 | 0.092981 |
| **D-Joining Bonus** | 0.120702 | 0.063783 | 2.790320e-02 | 0.039690 | -0.063680 | -0.046155 | 0.055443 | 0.048893 |
| **D-Agency** | 0.078631 | -0.007236 | 2.735501e-02 | 0.088777 | -0.177525 | 0.168445 | -0.052590 | -0.027735 |
| **D-Employee Referral** | -0.001364 | 0.026785 | -3.932174e-02 | -0.021252 | -0.040240 | -0.083977 | 0.086248 | 0.055677 |
| **Rex in Yrs** | -0.125276 | 0.067502 | 7.720122e-02 | 0.270473 | -0.391037 | -0.388453 | 0.397997 | 0.415723 |
| **D-Sales** | 0.013652 | 0.039892 | -5.716563e-20 | -0.014393 | -0.018520 | -0.049882 | 0.015605 | 0.047952 |
| **D-Healthcare** | 0.024335 | -0.005288 | 1.183410e-02 | 0.029144 | -0.036430 | 0.050842 | -0.028094 | -0.007891 |

| | Candidate Ref | D-DOJ Extended | Duration to accept offer | Notice period | E0 | E1 | E2 | E3 |
|---|---|---|---|---|---|---|---|---|
| **D-Infra** | -0.202255 | -0.101930 | -1.530722e-01 | -0.072286 | 0.001274 | -0.319310 | 0.250114 | 0.178546 |
| **D-AXON** | 0.034171 | 0.054859 | 5.332150e-02 | 0.060424 | -0.066344 | 0.026325 | 0.027380 | -0.023327 |
| **D-Bangalore** | -0.027485 | 0.020793 | 8.558013e-02 | 0.083707 | -0.150681 | 0.104944 | -0.010159 | -0.023634 |
| **D-Chennai** | 0.097617 | -0.020954 | 1.598392e-02 | -0.030952 | 0.035298 | 0.070796 | -0.067177 | -0.052566 |
| **D-Mumbai** | 0.027952 | -0.008478 | -4.758102e-02 | -0.036146 | -0.044152 | -0.002761 | 0.032699 | -0.005104 |
| **D-Delhi NCR** | -0.114463 | -0.009262 | -8.897154e-02 | -0.014253 | 0.065442 | -0.132598 | 0.064184 | 0.068464 |
| **D-Hyderabad** | 0.022961 | 0.040106 | 2.493119e-02 | 0.002396 | 0.022906 | -0.051586 | 0.029438 | 0.026254 |
| **Age** | 0.355756 | 0.046028 | 9.010494e-03 | -0.071338 | 0.027592 | -0.335565 | 0.192380 | 0.235454 |
| **D-Joining** | 0.056303 | 0.146467 | -4.985932e-02 | -0.244325 | 0.084974 | -0.018465 | -0.032048 | -0.001661 |

In [63]:
```
pd.crosstab(index=data['Age'],columns=data['Gender'])
```

Out[63]:

| Gender | Female | Male |
|---|---|---|
| **Age** | | |
| **19** | 0 | 2 |
| **20** | 1 | 1 |
| **21** | 0 | 5 |
| **22** | 42 | 43 |
| **23** | 139 | 167 |
| **24** | 132 | 353 |
| **25** | 178 | 646 |
| **26** | 220 | 835 |
| **27** | 232 | 920 |
| **28** | 161 | 847 |
| **29** | 142 | 724 |
| **30** | 102 | 639 |
| **31** | 73 | 544 |
| **32** | 48 | 426 |
| **33** | 40 | 328 |
| **34** | 700 | 2772 |
| **35** | 22 | 206 |
| **36** | 14 | 156 |
| **37** | 10 | 112 |
| **38** | 10 | 85 |
| **39** | 6 | 59 |
| **40** | 4 | 55 |
| **41** | 2 | 32 |
| **42** | 2 | 28 |
| **43** | 3 | 13 |
| **44** | 0 | 8 |
| **45** | 0 | 12 |
| **46** | 1 | 6 |
| **47** | 1 | 4 |
| **48** | 1 | 4 |
| **49** | 0 | 4 |
| **50** | 0 | 5 |

| Gender | Female | Male |
|---|---|---|
| Age | | |
| **51** | 0 | 3 |
| **60** | 0 | 1 |
| **62** | 0 | 1 |
| **64** | 0 | 1 |