In [1]:
```
import pandas as pd
from matplotlib import pyplot as plt
%matplotlib inline
```

```
Matplotlib is building the font cache; this may take a moment.
```

In [2]:
```
df = pd.read_csv("HR_comma_sep.csv")
df.head()
```

Out[2]:

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	Department	s
0	0.38	0.53	2	157	3	0	1	0	sales	
1	0.80	0.86	5	262	6	0	1	0	sales	me
2	0.11	0.88	7	272	4	0	1	0	sales	me
3	0.72	0.87	5	223	5	0	1	0	sales	
4	0.37	0.52	2	159	3	0	1	0	sales	

In [3]:
```
#Data Exploration and visualization
```

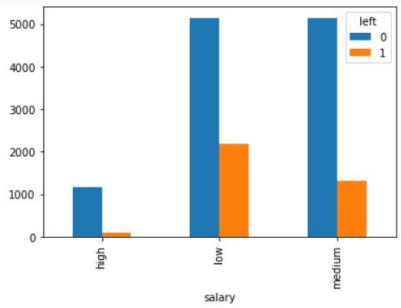
In [4]:
```
left = df[df.left==1]
left.shape
```

Out[4]:
```
(3571, 10)
```

In [6]:
```
retained = df[df.left==0]
retained.shape
```

Out[6]:
```
(11428, 10)
```

In [7]:
```
# Now we need to find out the parameters that affect the retention of an employee
```



In [14]:
```
# people with high salary are less likely to leave the firm
```

In [15]:
```
# From the data analysis so far we can conclude that we will use following variables as independant variables in our model
# 1**Satisfaction Level**
# 2**Average Monthly Hours**
# 3**Promotion Last 5 Years**
# 4**Salary**
```

In [16]:
```
# Tackle salary dummy variable

# Salary has all text data. It needs to be converted to numbers and we will use dummy variable for that.
# Check my one hot encoding tutorial to understand purpose behind dummy variables.
```

In [18]:
```
subdf = df[['satisfaction_level','average_montly_hours','promotion_last_5years','salary']]
subdf.head()
salary_dummies = pd.get_dummies(subdf.salary, prefix="salary")
df_with_dummies = pd.concat([subdf,salary_dummies],axis='columns')
df_with_dummies.head()
```

In [8]:
```
df.groupby('left').mean()
```

Out[8]:

left	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years
0	0.666810	0.715473	3.786664	199.060203	3.380032	0.175009	0.026251
1	0.440098	0.718113	3.855503	207.419210	3.876505	0.047326	0.005321

In [9]:
```
# Observe the data and explain
```

In [10]:
```

```

In [11]:
```
# **Satisfaction Level**: Satisfaction level seems to be relatively low (0.44) in employees leaving the firm vs the retained ones
# **Average Monthly Hours**: Average monthly hours are higher in employees leaving the firm (199 vs 207)
# **Promotion Last 5 Years**: Employees who are given promotion are likely to be retained at firm
```

In [12]:
```
#Impact of salary on employee retention
```

In [13]:
```
pd.crosstab(df.salary,df.left).plot(kind='bar')
```

Out[13]:
```
<AxesSubplot:xlabel='salary'>
```

Out[18]:

	satisfaction_level	average_montly_hours	promotion_last_5years	salary	salary_high	salary_low	salary_medium
0	0.38	157	0	low	0	1	0
1	0.80	262	0	medium	0	0	1
2	0.11	272	0	medium	0	0	1
3	0.72	223	0	low	0	1	0
4	0.37	159	0	low	0	1	0

In [19]:
```
#Now we need to remove salary column which is text data. It is already replaced by dummy variables so we can safely remove it
```

In [20]:
```
df_with_dummies.drop('salary',axis='columns',inplace=True)
df_with_dummies.head()
```

Out[20]:

	satisfaction_level	average_montly_hours	promotion_last_5years	salary_high	salary_low	salary_medium
0	0.38	157	0	0	1	0
1	0.80	262	0	0	0	1
2	0.11	272	0	0	0	1
3	0.72	223	0	0	1	0
4	0.37	159	0	0	1	0

In [21]:
```
X = df_with_dummies
X.head()
```

Out[21]:

	satisfaction_level	average_montly_hours	promotion_last_5years	salary_high	salary_low	salary_medium
0	0.38	157	0	0	1	0
1	0.80	262	0	0	0	1
2	0.11	272	0	0	0	1
3	0.72	223	0	0	1	0
4	0.37	159	0	0	1	0