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

1. (25%) Expand the Abstract from last homework, and add at least two additional problem statements that the dataset could be analyzed to answer them preliminarily, if not fully.

Aim is to answer an interesting question of a company such as "Why are our best and most experienced employees leaving prematurely? Continuing to investigate the same question by narrowing down to the focus areas, like which Department have high attrition, which department have more people with less salary and so on to find the root cause of the attrition prematurely (as we found out earlier that the mode number of 'time_spend_company' variable is 3.0 yrs)

For this exercise, we will use the dataset which was output of last week's exercise. The original dataset is formed by the Human Resources (HR) department after conducting a survey on their employees available at https://www.kaggle.com/cezarschroeder/human-resource-analytics-dataset Originally it had 14999 rows and 11 columns, however we dropped one column 'is_smoker', which is not relevant to our study. We manipulated for missing data and identified outliers based on 'average monthly hrs'. However, in this study, I have included that data because I need to investigate:

Q1. I want to know emplyee satisfaction level in every department, with their corresponding salary bracket and find out if employees with high salary were having high satisfaction level

Q2. What is the picture of satsifaction level among the employees those were promoted in last 5 yrs to those who were not, with their respective salary brackets

This investigation is not influenced by 'average_montly_hours'. Hence it contains all 1499 rows with 10 columns.

A copy of the dataset file is to be submitted alongside with the Jupyter Notebook report.

In [2]:
```
##Improting output data from Homework

hrdb3 = pd.read_csv("outputdata3.csv")
hrdb3.describe()
```

Out[2]:

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	work_accident	promotion_last_5years
count	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000	14999.000000
mean	0.612834	0.716102	3.803054	200.149743	3.489166	0.144610	0.021268
std	0.248631	0.171169	1.232592	49.647584	1.452451	0.351719	0.144281
min	0.090000	0.360000	2.000000	96.000000	2.000000	0.000000	0.000000
25%	0.440000	0.560000	3.000000	156.000000	3.000000	0.000000	0.000000
50%	0.640000	0.720000	4.000000	197.000000	3.000000	0.000000	0.000000
75%	0.820000	0.870000	5.000000	244.000000	4.000000	0.000000	0.000000
max	1.000000	1.000000	7.000000	310.000000	10.000000	1.000000	1.000000

In [3]:
```
hrdb3.shape
```

Out[3]: (14999, 10)

In [4]:
```
###Q1. I want to know emplyee satisfaction level in every department, with their corresponding salary bracket and
##find out if employees with high salary were having high satisfaction level

TDS_dept_sal = hrdb3.set_index(['department','salary'])
TDS_dept_sal_Sorted = TDS_dept_sal.sort_index()
TDS_dept_sal_Sorted = TDS_dept_sal_Sorted.loc[:, ('satisfaction_level', 'number_project', 'time_spend_company')]
TDS_dept_sal_Sorted
```

Out[4]:

		satisfaction_level	number_project	time_spend_company
department	**salary**			
IT	**high**	0.75	5	5
	high	0.46	2	3
	high	0.40	2	3
	high	0.72	5	5
	high	0.49	5	3

		satisfaction_level	number_project	time_spend_company
department	**salary**			
...	**...**	...	...	...
technical	**medium**	0.09	6	4
	medium	0.38	2	3
	medium	0.72	5	5
	medium	0.40	2	3

In [5]:
```
# Lets find out what is the average satisfaction level among all the employees in the dataset on the scale of 0 to 1,
# where is 1 is highly satisfied
val = TDS_dept_sal_Sorted['satisfaction_level'].mean().round(2)
print("The average satisfaction_level for all employees:", val)
```

```
The average satisfaction_level for all employees: 0.61
```

In [6]:
```
###Q2. What is the picture of satsifaction level among the employees those were promoted in last 5 yrs to those who w
##with their respective salary brackets

TDS_promoted_sal = hrdb3.set_index(['promotion_last_5years','salary'])
TDS_promoted_sal_Sorted = TDS_promoted_sal.sort_index()
TDS_promoted_sal_Sorted = TDS_promoted_sal_Sorted.loc[:, ('satisfaction_level', 'number_project','average_montly_hou
TDS_promoted_sal_Sorted
```

Out[6]:

		satisfaction_level	number_project	average_montly_hours
promotion_last_5years	**salary**			
0	**high**	0.45	2	149
	high	0.09	6	168
	high	0.44	2	156
	high	0.45	2	129
	high	0.37	2	149
...	**...**	...	...	...
1	**medium**	0.68	4	146
	medium	0.75	4	263
	medium	0.29	5	134
	medium	0.81	5	250
	medium	0.41	2	154

14999 rows × 3 columns

1. (25%) DataWrangling Playground.

In [7]:
```
## Creating a copy of a dataset as a recovery point.

hrdb4 = hrdb3.copy()
hrdb4.shape
```

Out[7]:
```
(14999, 10)
```

In [8]:
```
#Created pivot table using groupby() to find the average satisfaction level in each department and each salary bracke
subset_satis_mean = hrdb4.groupby(['department', 'salary']).satisfaction_level.mean()
subset_satis_mean
```

Out[8]:
```
department   salary
IT           high      0.638193
             low       0.610099
             medium    0.624187
RandD        high      0.586667
             low       0.623929
             medium    0.620349
accounting   high      0.614054
             low       0.574162
             medium    0.583642
hr           high      0.673111
             low       0.608657
             medium    0.580306
management   high      0.653333
```

```
             low       0.610722
             medium    0.597867
marketing    high      0.605250
             low       0.602910
             medium    0.638218
product_mng  high      0.614118
             low       0.620909
             medium    0.619112
sales        high      0.648959
             low       0.600838
             medium    0.625327
support      high      0.655035
             low       0.591710
             medium    0.645149
technical    high      0.625970
             low       0.594322
             medium    0.620968
Name: satisfaction_level, dtype: float64
```

In [9]:
```
subset_satis_mean.index
```

Out[9]:
```
MultiIndex([(         'IT',   'high'),
            (         'IT',    'low'),
            (         'IT', 'medium'),
            (      'RandD',   'high'),
            (      'RandD',    'low'),
            (      'RandD', 'medium'),
            ( 'accounting',   'high'),
            ( 'accounting',    'low'),
            ( 'accounting', 'medium'),
            (         'hr',   'high'),
            (         'hr',    'low'),
            (         'hr', 'medium'),
            ( 'management',   'high'),
            ( 'management',    'low'),
            ( 'management', 'medium'),
            (  'marketing',   'high'),
            (  'marketing',    'low'),
            (  'marketing', 'medium'),
            ('product_mng',   'high'),
            ('product_mng',    'low'),
            ('product_mng', 'medium'),
            (      'sales',   'high'),
            (      'sales',    'low'),
            (      'sales', 'medium'),
            (    'support',   'high'),
            (    'support',    'low'),
            (    'support', 'medium'),
            (  'technical',   'high'),
            (  'technical',    'low'),
            (  'technical', 'medium')],
           names=['department', 'salary'])
```

In [10]:
```
##Converting to indexed dataframe using unstack()

subset_unstack = subset_satis_mean.unstack()
subset_unstack
```

Out[10]:

salary	high	low	medium
department			
IT	0.638193	0.610099	0.624187
RandD	0.586667	0.623929	0.620349
accounting	0.614054	0.574162	0.583642
hr	0.673111	0.608657	0.580306
management	0.653333	0.610722	0.597867
marketing	0.605250	0.602910	0.638218
product_mng	0.614118	0.620909	0.619112
sales	0.648959	0.600838	0.625327
support	0.655035	0.591710	0.645149
technical	0.625970	0.594322	0.620968

In [11]:
```
# Identifying if melt() gives any valuable information

melted = pd.melt(hrdb4, ['department', 'salary'])
melted.head(10)
```

Out[11]:

	department	salary	variable	value
0	sales	low	satisfaction_level	0.38
1	sales	medium	satisfaction_level	0.8
2	sales	medium	satisfaction_level	0.11
3	sales	low	satisfaction_level	0.72
4	sales	low	satisfaction_level	0.37
5	sales	low	satisfaction_level	0.41
6	sales	low	satisfaction_level	0.1
7	sales	low	satisfaction_level	0.92
8	sales	low	satisfaction_level	0.89
9	sales	low	satisfaction_level	0.42

In [12]:
```
##Creating pivot table to see what are the average years spent by an employee before leaving, by deppartment and by s

dataframe_reset = TDS_dept_sal_Sorted.reset_index()

pivoted = np.round(hrdb4.pivot_table(index='salary', columns='department', values='time_spend_company', aggfunc='mean
pivoted
```

Out[12]:

department	IT	RandD	accounting	hr	management	marketing	product_mng	sales	support	technical
salary										
high	3.07	3.53	3.22	2.91	5.16	3.50	3.62	3.51	3.20	3.31
low	3.43	3.38	3.42	3.25	3.41	3.51	3.42	3.46	3.47	3.39
medium	3.57	3.31	3.68	3.49	4.07	3.61	3.49	3.61	3.30	3.44

In [13]:
```
dataframe_reset2 = TDS_promoted_sal_Sorted.reset_index()
dataframe_reset2.head(10)
```

Out[13]:

	promotion_last_5years	salary	satisfaction_level	number_project	average_montly_hours
0	0	high	0.45	2	149
1	0	high	0.09	6	168
2	0	high	0.44	2	156
3	0	high	0.45	2	129
4	0	high	0.37	2	149
5	0	high	0.10	6	278
6	0	high	0.36	2	156
7	0	high	0.40	2	143
8	0	high	0.80	3	255
9	0	high	0.66	5	161

In [14]:
```
##Creating multi-index pivot table to see what were the average number of projects worked by an employee before leavi
## was s/he prpmoted in last 5 yrs, by salary bracket and by satisfaction level

dataframe_reset2.pivot_table(index=['salary', 'satisfaction_level'],columns= ['promotion_last_5years'], values= 'numb
```

Out[14]:

promotion_last_5years		0	1
salary	satisfaction_level		
high	0.09	6.000000	NaN
	0.10	6.111111	NaN
	0.11	6.500000	NaN
	0.12	6.000000	NaN
	0.13	4.250000	NaN

promotion_last_5years		0	1
salary	satisfaction_level		
...	...	...	...
medium	**0.96**	3.772727	4.0
	0.97	3.745763	4.4
	0.98	3.763441	NaN
	0.99	3.829268	NaN
	1.00	3.903846	6.0

1. (25%) Summary and Conclusion

The employees of HR department with High salary bracket were the happiest employees (0.67) who left the organization. And the employees of Accounts department with Low salary bracket were the dissatisfied employees (0.57) who left the organization.

The employees of Technical department with High salary bracket were happy with average satisfaction level of 0.86, who were promoted once is last 5 years, however, the employees with Accounts department with low salary backet were dissatisfied with average level of 0.57, who were not promoted in last 5 years.

We also, see from the pivot table, employees of HR department inspite of having high salary leave early (in average of 2.9 yrs), However, employees of Management department stays longer with company (avg 5.16 yrs) when they were drawing high salary.

These conclusions are based on the average satisfaction level. More study can be done using visualizations and problem area can be further narrowed down.