In [99]:
```
import pandas as pd# we import pandas to handle the file and its save our lots of time, it handle large data set efficeintly.
import numpy as np #for creating arrays
import seaborn as sns #for data visualzation
import matplotlib.pyplot as plt # for data visualization
```

In [100…
```
df1 = pd.read_csv(r'C:\Users\PC-chetan\Desktop\train.csv') # trian data
```

In [101…
```
df2 = pd.read_csv(r'C:\Users\PC-chetan\Desktop\test.csv') # test data
```

In [102…
```
# here two data sets are given
# train data and the test data
# in test data column is_ promoted is missing because the data is alredy split in the two parts
# and is_proomoted is the column of which values we have to pridict
```

Step 4 Missing Values Treatment

- Reasons of having Missing Values in Data set,
- Unavailability of Data
- Loss of Data
- Data Entry Error
- Incomplete Forms etc.

- Dealing with Null values of each columns and also with other inappropriate data in the column

In [106…
```
df1.isnull().sum()
```

Out[106…
```
employee_id               0
department                0
region                    0
education              2409
gender                    0
recruitment_channel       0
no_of_trainings           0
age                       0
previous_year_rating    4124
length_of_service         0
awards_won?               0
avg_training_score        0
is_promoted               0
dtype: int64
```

In [107…
```
# luckly we have only two clumn in each data which we havbe to deal with
```

In [108…
```
df2.isnull().sum()
```

Out[108…
```
employee_id               0
department                0
region                    0
education              1034
gender                    0
recruitment_channel       0
no_of_trainings           0
age                       0
previous_year_rating    1812
length_of_service         0
awards_won?               0
avg_training_score        0
dtype: int64
```

In [109…
```
df1.education.value_counts()
```

Out[109…
```
Bachelor's          36669
Master's & above    14925
Below Secondary       805
Name: education, dtype: int64
```

In [110…
```
df1.education.unique()
```

Out[110…
```
array(["Master's & above", "Bachelor's", nan, 'Below Secondary'],
      dtype=object)
```

In [111…
```
df1.education.describe()
```

Out[111…
```
count          52399
unique             3
top       Bachelor's
freq           36669
Name: education, dtype: object
```

In [112…
```
# now we will fill the null values with the most frequent value of the column
```

In [113…
```
df1.education.fillna("Bachelor's", inplace=True)
```

In [114…
```
df2.education.fillna("Bachelor's", inplace=True)
```

In [115…
```
df1.education.isnull().sum()
```

Out[115…
```
0
```

In [116…
```
df2.education.isnull().sum()
```

Out[116…
```
0
```

In [117…
```
df1.previous_year_rating.value_counts()
```

Out[117…
```
3.0    18618
5.0    11741
4.0     9877
1.0     6223
2.0     4225
Name: previous_year_rating, dtype: int64
```

In [118…
```
df1.previous_year_rating.unique()
```

Out[118…
```
array([ 5.,  3.,  1.,  4., nan,  2.])
```

In [119…
```
df1.head()
```

Out[119…

	employee_id	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	awards_won?	avg_training_score	is_promoted
0	65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5.0	8	0	49	0
1	65141	Operations	region_22	Bachelor's	m	other	1	30	5.0	4	0	60	0
2	7513	Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34	3.0	7	0	50	0
3	2542	Sales & Marketing	region_23	Bachelor's	m	other	2	39	1.0	10	0	50	0
4	48945	Technology	region_26	Bachelor's	m	other	1	45	3.0	2	0	73	0

In [120…
```
df1.previous_year_rating.fillna('3.0',inplace=True)
```

In [121…
```
df2.previous_year_rating.fillna('3.0',inplace=True)
```

In [122…
```
# now we have sucessfully clean our data
```

In [123…
```
df1.isnull().sum()
```

Out[123…
```
employee_id             0
department              0
region                  0
education               0
gender                  0
recruitment_channel     0
no_of_trainings         0
age                     0
previous_year_rating    0
length_of_service       0
awards_won?             0
avg_training_score      0
is_promoted             0
dtype: int64
```

In [124…
```
df2.isnull().sum()
```

Out[124…
```
employee_id             0
department              0
region                  0
education               0
gender                  0
recruitment_channel     0
no_of_trainings         0
age                     0
previous_year_rating    0
length_of_service       0
awards_won?             0
avg_training_score      0
dtype: int64
```

In [ ]:
```
```