In [ ]:
```
# Practicle 1 : Preprocessing of dataset
# Preprocessing for data sciencentist salary dataset
# 2021BIT023
```

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
```

In [2]:
```
df=pd.read_csv(ds_salaries1.csv)
df
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[2], line 1
----> 1 df=pd.read_csv(ds_salaries1.csv)
      2 df

NameError: name 'ds_salaries1' is not defined
```

In [5]:
```
df=pd.read_csv("ds_salaries1.csv")
```

In [6]:
```
df.head(5)
```

Out[6]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	e
0	2023	SE	FT	Principal Data Scientist	80000.0	EUR	85847.0	
1	2023	MI	CT	ML Engineer	30000.0	USD	30000.0	
2	2023	MI	CT	ML Engineer	25500.0	USD	25500.0	
3	2023	SE	FT	Data Scientist	175000.0	USD	175000.0	
4	2023	SE	FT	Data Scientist	120000.0	USD	120000.0	

In [7]:
```
df.tail(5)
```

Out[7]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_u
3750	2020	SE	FT	Data Scientist	412000.0	USD	41200(
3751	2021	MI	FT	Principal Data Scientist	151000.0	USD	15100(
3752	2020	EN	FT	Data Scientist	105000.0	USD	10500(
3753	2020	EN	CT	Business Data Analyst	100000.0	USD	10000(
3754	2021	SE	FT	Data Science Manager	7000000.0	INR	9466!

In [8]:
```
df.shape # row, attributes
```

Out[8]:
```
(3755, 11)
```

In [9]:
```
df1=df.copy()
```

In [10]:
```
df1.head(2)
```

Out[10]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	er
0	2023	SE	FT	Principal Data Scientist	80000.0	EUR	85847.0	
1	2023	MI	CT	ML Engineer	30000.0	USD	30000.0	

In [9]:
```
df1.isnull().sum()
```

Out[9]:
```
work_year            0
experience_level     0
employment_type      0
job_title            0
salary               2
salary_currency      1
salary_in_usd        3
employee_residence   0
remote_ratio         1
company_location     1
company_size         0
dtype: int64
```

In [28]:
```
# vool series
print(df1.isna()) # false for not null data aani true for NAN value i.e for null value
```

```
      work_year  experience_level  employment_type  job_title  salary  \
0         False             False            False      False   False
1         False             False            False      False   False
2         False             False            False      False   False
3         False             False            False      False   False
4         False             False            False      False   False
...         ...               ...              ...        ...     ...
3750      False             False            False      False   False
3751      False             False            False      False   False
3752      False             False            False      False   False
3753      False             False            False      False   False
3754      False             False            False      False   False

      salary_currency  salary_in_usd  employee_residence  remote_ratio  \
0               False          False               False         False
1               False          False               False         False
2               False          False               False         False
3               False          False               False         False
4               False          False               False         False
...               ...            ...                 ...           ...
3750            False          False               False         False
3751            False          False               False         False
3752            False          False               False         False
3753            False          False               False         False
3754            False          False               False         False

      company_location  company_size
0                False         False
1                False         False
2                False         False
3                False         False
4                False         False
...                ...           ...
3750             False         False
3751             False         False
3752             False         False
3753             False         False
3754             False         False

[3751 rows x 11 columns]
```

In [12]:
```
df1.dropna(inplace=True)
```

In [13]:
```
df1.isnull().sum()
```

Out[13]:
```
work_year            0
experience_level     0
employment_type      0
job_title            0
salary               0
salary_currency      0
salary_in_usd        0
employee_residence   0
remote_ratio         0
company_location     0
company_size         0
dtype: int64
```

In [15]:
```
# in salary_currency we replce salary from INR to USD
df_repl=df1.replace({'INR':'USD'})
```

```
df_repl.head(3)
```

Out[15]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	er
0	2023	SE	FT	Principal Data Scientist	80000.0	EUR	85847.0	
1	2023	MI	CT	ML Engineer	30000.0	USD	30000.0	
2	2023	MI	CT	ML Engineer	25500.0	USD	25500.0	

In [16]:
```
df_intrplt=df1.interpolate()
df_intrplt.head(2)
```

Out[16]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	er
0	2023	SE	FT	Principal Data Scientist	80000.0	EUR	85847.0	
1	2023	MI	CT	ML Engineer	30000.0	USD	30000.0	

In [17]:
```
df1=df1.infer_object()
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_6456\2477449844.py in ?()
----> 1 df1=df1.infer_object()

~\anaconda3\Lib\site-packages\pandas\core\generic.py in ?(self, name)
   5985             and name not in self._accessors
   5986             and self._info_axis._can_hold_identifiers_and_holds_name(name)
   5987         ):
   5988             return self[name]
-> 5989         return object.__getattribute__(self, name)

AttributeError: 'DataFrame' object has no attribute 'infer_object'
```

In [19]:
```
df2=df1.infer_objects()
```

In [20]:
```
df2
```

Out[20]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_u
0	2023	SE	FT	Principal Data Scientist	80000.0	EUR	8584
1	2023	MI	CT	ML Engineer	30000.0	USD	3000
2	2023	MI	CT	ML Engineer	25500.0	USD	2550
3	2023	SE	FT	Data Scientist	175000.0	USD	17500
4	2023	SE	FT	Data Scientist	120000.0	USD	12000
...	...	...	...	...	...	...	
3750	2020	SE	FT	Data Scientist	412000.0	USD	41200
3751	2021	MI	FT	Principal Data Scientist	151000.0	USD	15100
3752	2020	EN	FT	Data Scientist	105000.0	USD	10500
3753	2020	EN	CT	Business Data Analyst	100000.0	USD	10000
3754	2021	SE	FT	Data Science Manager	7000000.0	INR	9466

3751 rows × 11 columns

In [18]:
```
df1=df1.interpolate()
```

In [19]:
```
print(df1.notnull())  # true for not missing value , false for missing value
```

```
      work_year  experience_level  employment_type  job_title  salary  \
0          True              True             True       True    True
1          True              True             True       True    True
2          True              True             True       True    True
3          True              True             True       True    True
4          True              True             True       True    True
...         ...               ...              ...        ...     ...
3750       True              True             True       True    True
3751       True              True             True       True    True
3752       True              True             True       True    True
3753       True              True             True       True    True
3754       True              True             True       True    True

      salary_currency  salary_in_usd  employee_residence  remote_ratio  \
0                True           True                True          True
1                True           True                True          True
2                True           True                True          True
3                True           True                True          True
4                True           True                True          True
...               ...            ...                 ...           ...
3750             True           True                True          True
3751             True           True                True          True
3752             True           True                True          True
3753             True           True                True          True
3754             True           True                True          True

      company_location  company_size
0                 True          True
1                 True          True
2                 True          True
3                 True          True
4                 True          True
...                ...           ...
3750              True          True
3751              True          True
3752              True          True
3753              True          True
3754              True          True

[3751 rows x 11 columns]
```

In [20]:
```
df1.notnull().sum()
```

Out[20]:
```
work_year             3751
experience_level      3751
employment_type       3751
job_title             3751
salary                3751
salary_currency       3751
salary_in_usd         3751
employee_residence    3751
remote_ratio          3751
company_location      3751
company_size          3751
dtype: int64
```

In [21]:
```
df2=df.copy()
```

In [23]:
```
df2_filna=df2.fillna(0)
```

In [24]:
```
#filter data
filter_data=df2[df['salary']>=81000]
filter_data
```

Out[24]:

	work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_u
3	2023	SE	FT	Data Scientist	175000.0	USD	17500(
4	2023	SE	FT	Data Scientist	120000.0	USD	12000(
5	2023	SE	FT	Applied Scientist	222200.0	USD	22220(
6	2023	SE	FT	Applied Scientist	136000.0	USD	13600(
7	2023	SE	FT	Data Scientist	219000.0	USD	21900(
...	...	...	...	...	...	...	
3750	2020	SE	FT	Data Scientist	412000.0	USD	41200(
3751	2021	MI	FT	Principal Data Scientist	151000.0	USD	15100(
3752	2020	EN	FT	Data Scientist	105000.0	USD	10500(
3753	2020	EN	CT	Business Data Analyst	100000.0	USD	10000(
3754	2021	SE	FT	Data Science Manager	7000000.0	INR	9466[

3082 rows × 11 columns

In [37]:
```
# dict2lst=df2.to_dict(orient='list')
# print(dict2lst)
```