```
In [19]:
          import pandas as pd
           # Correct method to read an Excel file
           emp = pd.read_excel('/content/drive/MyDrive/Rawdata.xlsx')
           print(emp)
                Name
                                Domain
                                               Age
                                                      Location
                                                                  Salary
                                                                                Exp
          0
                Mike
                        Datascience#$
                                         34 years
                                                        Mumbai
                                                                  5^00#0
                                                                                 2+
          1
             Teddy^
                               Testing
                                           45' yr
                                                                 10%%000
                                                     Bangalore
                                                                                 <3
          2
               Uma#r
                       Dataanalyst^^#
                                               NaN
                                                                 1$5%000
                                                                             4> yrs
                                                           NaN
          3
                Jane
                          Ana^^lytics
                                               NaN
                                                      Hyderbad
                                                                  2000^0
                                                                                NaN
          4
             Uttam*
                           Statistics
                                             67-yr
                                                           NaN
                                                                  30000-
                                                                           5+ year
          5
                 Kim
                                   NLP
                                              55yr
                                                         Delhi
                                                                 6000^$0
                                                                                10+
In [20]:
           emp.columns
          Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
Out[20]:
In [21]:
           emp.shape
           (6, 6)
Out[21]:
In [22]:
           emp.head()
                          Domain
                                            Location
              Name
                                      Age
                                                        Salary
                                                                 Exp
Out[22]:
                                  34 years
                                                                   2+
               Mike
                     Datascience#$
                                             Mumbai
                                                       5^00#0
           1 Teddy^
                           Testing
                                     45' yr
                                           Bangalore
                                                     10%%000
                                                                   <3
             Uma#r
                    Dataanalyst^^#
                                      NaN
                                               NaN
                                                     1$5%000
                                                                4> yrs
               Jane
                        Ana^^lytics
                                      NaN
                                           Hyderbad
                                                       2000^0
                                                                 NaN
             Uttam*
                         Statistics
                                               NaN
                                                              5+ year
                                     67-yr
                                                       30000-
In [23]:
           emp.tail()
              Name
                          Domain
                                         Location
Out[23]:
                                   Age
                                                     Salary
                                                               Exp
                                  45' yr
           1 Teddy^
                                        Bangalore
                                                  10%%000
                                                                <3
                           Testing
             Uma#r
                    Dataanalyst^^#
                                   NaN
                                             NaN
                                                   1$5%000
                                                             4> yrs
                        Ana^^lytics
                                   NaN
                                         Hyderbad
                                                    2000^0
               Jane
                                                               NaN
             Uttam*
                                                     30000-
                                                            5+ year
                          Statistics
                                  67-yr
                                             NaN
          5
                Kim
                             NLP
                                   55yr
                                            Delhi
                                                    6000^$0
                                                               10+
In [24]:
           emp.info()
          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 6 entries, 0 to 5
          Data columns (total 6 columns):
           #
                Column
                           Non-Null Count
                                              Dtype
           - - -
           0
                Name
                           6 non-null
                                              object
           1
                           6 non-null
                Domain
                                              object
           2
                           4 non-null
                                              object
                Age
           3
                Location 4 non-null
                                              object
```

5

Salary

Exp

6 non-null

5 non-null

object

object

dtypes: object(6)

memory usage: 416.0+ bytes

In [25]: emp

Out[25]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience#\$	34 years	Mumbai	5^00#0	2+
1	Teddy^	Testing	45' yr	Bangalore	10%%000	<3
2	Uma#r	Dataanalyst^^#	NaN	NaN	1\$5%000	4> yrs
3	Jane	Ana^^lytics	NaN	Hyderbad	2000^0	NaN
4	Uttam*	Statistics	67-yr	NaN	30000-	5+ year
5	Kim	NLP	55yr	Delhi	6000^\$0	10+

```
In [26]: emp['Domain']
```

Out[26]:

Domain

- **0** Datascience#\$
- 1 Testing
- 2 Dataanalyst^^#
- 3 Ana^^lytics
- 4 Statistics
- 5 NLP

dtype: object

In [11]: en

emp.isnull()

Out	[11]
Uul	44

Name	Domain	Age	Location	Salary	Exp
False	False	False	False	False	False
False	False	False	False	False	False
False	False	True	True	False	False
False	False	True	False	False	True
False	False	False	True	False	False
False	False	False	False	False	False
	False False False False False	False	False False False False False False False False True False False True False False False	False False False False False False False False False True True False False True False False False False True	FalseFalseFalseFalseFalseFalseTrueTrueFalseFalseFalseTrueFalseFalseFalseFalseFalseTrueFalse

Out[27]:

0

Name 0

Domain 0

Age 2

Location 2

Salary 0

Exp 1

```
dtype: int64
In [28]:
          emp['Name']
Out[28]:
              Name
              Mike
          1 Teddy^
             Uma#r
               Jane
          4 Uttam*
                Kim
         dtype: object
          emp['Name'] = emp['Name'].str.replace(r'\W','',regex=True)
In [32]:
          emp['Name']
In [33]:
Out[33]:
             Name
             Mike
          1 Teddy
             Umar
              Jane
             Uttam
               Kim
         dtype: object
          emp['Domain']
In [34]:
Out[34]:
                  Domain
          0 Datascience#$
                   Testing
          2 Dataanalyst^^#
          3
                Ana^^lytics
          4
                  Statistics
                     NLP
         dtype: object
          emp['Domain'] = emp['Domain'].str.replace(r'\W','',regex=True)
In [35]:
In [36]:
          emp['Domain']
Out[36]:
                Domain
```

```
4
                  Statistics
                     NLP
           5
          dtype: object
In [37]:
           emp
Out[37]:
               Name
                          Domain
                                       Age
                                             Location
                                                          Salary
                                                                     Exp
           0
                Mike
                      Datascience
                                   34 years
                                              Mumbai
                                                          5^00#0
                                                                      2+
               Teddy
                                      45' yr
                                            Bangalore
                                                       10%%000
                                                                      <3
                          Testing
           2
               Umar
                      Dataanalyst
                                      NaN
                                                 NaN
                                                        1$5%000
                                                                   4> yrs
                                                          2000^0
           3
                Jane
                         Analytics
                                      NaN
                                             Hyderbad
                                                                     NaN
                                                                  5+ year
               Uttam
                         Statistics
                                      67-yr
                                                 NaN
                                                          30000-
           5
                Kim
                             NLP
                                      55yr
                                                 Delhi
                                                         6000^$0
                                                                     10+
In [38]:
            emp['Age']=emp['Age'].str.replace(r'\D','',regex=True)
In [40]:
            emp['Age']
Out[40]:
               Age
           0
                34
            1
                45
               NaN
           3
               NaN
                67
            5
                55
          dtype: object
In [41]:
            emp
Out[41]:
               Name
                          Domain
                                   Age
                                         Location
                                                      Salary
                                                                 Exp
           0
                                    34
                                                                   2+
                Mike
                      Datascience
                                           Mumbai
                                                      5^00#0
               Teddy
                           Testing
                                    45
                                         Bangalore
                                                   10%%000
                                                                   <3
           2
               Umar
                      Dataanalyst
                                   NaN
                                              NaN
                                                    1$5%000
                                                                4> yrs
                                         Hyderbad
                                                      2000^0
                                                                 NaN
                Jane
                         Analytics
                                   NaN
                         Statistics
                                                      30000-
                                                              5+ year
               Uttam
                                    67
                                              NaN
                             NLP
                                    55
                                             Delhi
                                                     6000^$0
            5
                 Kim
                                                                 10+
```

emp['Location']=emp['Location'].str.replace(r'\W','',regex=True)

Datascience

Dataanalyst

Analytics

Testing

1

2

3

In [42]:

```
In [43]:
           emp['Location']
Out[43]:
              Location
               Mumbai
             Bangalore
           2
                  NaN
              Hyderbad
                  NaN
                  Delhi
          dtype: object
           emp['Salary']
In [44]:
Out[44]:
                Salary
                5^00#0
           1 10%%000
              1$5%000
                2000^0
          3
           4
                30000-
               6000^$0
          dtype: object
          emp['Salary']=emp['Salary'].str.replace(r'\D','',regex=True)
In [46]:
In [47]:
           emp['Salary']
Out[47]:
              Salary
               5000
              10000
              15000
              20000
              30000
              60000
          dtype: object
In [48]:
Out[48]:
             Name
                       Domain
                                Age
                                      Location Salary
                                                         Exp
              Mike
                    Datascience
                                 34
                                       Mumbai
                                                 5000
                                                          2+
                                                10000
             Teddy
                        Testing
                                 45
                                     Bangalore
                                                           <3
```

4> yrs

NaN

Umar

Dataanalyst

NaN

```
emp['Exp']=emp['Exp'].str.replace(r'\D','',regex=True)
In [49]:
           emp['Exp']
In [50]:
Out[50]:
               Exp
                 2
           0
                 3
           2
                 4
              NaN
                 5
                10
          dtype: object
In [51]:
           emp
Out[51]:
              Name
                         Domain
                                  Age
                                        Location Salary
                                                          Exp
               Mike
                     Datascience
                                   34
                                         Mumbai
                                                   5000
                                                            2
                                                  10000
                                                            3
              Teddy
                         Testing
                                   45
                                       Bangalore
                                                  15000
                                                            4
               Umar
                      Dataanalyst
                                 NaN
                                            NaN
           3
               Jane
                        Analytics
                                 NaN
                                        Hyderbad
                                                  20000
                                                         NaN
                        Statistics
                                   67
                                                  30000
                                                            5
              Uttam
                                            NaN
                                   55
                                                  60000
                Kim
                            NLP
                                            Delhi
                                                           10
In [52]:
           clean_data=emp.copy()
In [53]:
           clean_data
Out[53]:
              Name
                         Domain
                                  Age
                                        Location Salary
                                                          Exp
           0
               Mike
                     Datascience
                                   34
                                         Mumbai
                                                   5000
                                                            2
                                                  10000
                                                            3
              Teddy
                          Testing
                                   45
                                       Bangalore
                                 NaN
               Umar
                      Dataanalyst
                                            NaN
                                                  15000
                                                            4
               Jane
                        Analytics
                                 NaN
                                        Hyderbad
                                                  20000
                                                         NaN
                                                  30000
                                                            5
           4
              Uttam
                        Statistics
                                   67
                                            NaN
                            NLP
                                                  60000
                Kim
                                   55
                                            Delhi
                                                           10
           clean_data['Age']
In [54]:
Out[54]:
               Age
```

Hyderbad

NaN

Delhi

Jane

Uttam

Kim

5

0

34

Analytics

Statistics

NLP

NaN

67

55

20000

30000

60000

NaN

10+

5+ year

```
55
         dtype: object
In [55]:
          import numpy as np
          clean_data['Age']=clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age'])))
In [56]:
In [57]:
          clean_data['Age']
Out[57]:
              Age
          0
               34
               45
          2 50.25
          3 50.25
               67
               55
         dtype: object
In [58]:
          clean_data['Exp']
Out[58]:
             Exp
               2
          0
               3
          2
               4
          3 NaN
               5
              10
         dtype: object
          clean_data['Exp']=clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp'])))
In [59]:
In [60]:
          clean_data['Exp']
Out[60]:
            Exp
               2
               3
               4
             4.8
```

67

2 NaN3 NaN

```
dtype: object
In [61]:
           clean_data
Out[61]:
              Name
                        Domain
                                  Age
                                        Location
                                                 Salary Exp
                                                           2
                    Datascience
                                   34
                                                   5000
               Mike
                                         Mumbai
             Teddy
                                   45
                                       Bangalore
                                                  10000
                                                           3
                         Testing
           2
                                50.25
                                                  15000
                                                           4
              Umar
                     Dataanalyst
                                            NaN
                                50.25
                                       Hyderbad
                                                  20000
               Jane
                       Analytics
                                                          4.8
              Uttam
                       Statistics
                                   67
                                            NaN
                                                  30000
                                                           5
           5
               Kim
                           NLP
                                   55
                                           Delhi
                                                  60000
                                                          10
           clean_data['Location'].isnull().sum()
In [62]:
Out[62]:
In [63]:
           clean_data['Location']
Out[63]:
              Location
           0
               Mumbai
             Bangalore
           2
                  NaN
              Hyderbad
                  NaN
           5
                  Delhi
          dtype: object
           clean_data['Location']=clean_data['Location'].fillna(clean_data['Location'].mode()[0])
In [64]:
In [65]:
           clean_data['Location']
Out[65]:
              Location
           0
               Mumbai
             Bangalore
             Bangalore
              Hyderbad
             Bangalore
           5
                  Delhi
          dtype: object
```

```
Out[66]:
            Name
                      Domain
                               Age
                                    Location Salary Exp
          0
             Mike
                  Datascience
                                34
                                     Mumbai
                                              5000
                                                     2
          1 Teddy
                      Testing
                                45
                                   Bangalore
                                             10000
                                                     3
                   Dataanalyst
                             50.25
                                   Bangalore
                                             15000
                                                     4
             Umar
          3
                             50.25
                                             20000
             Jane
                     Analytics
                                   Hyderbad
                                                    4.8
            Uttam
                     Statistics
                                67
                                   Bangalore
                                             30000
                                                     5
          5
              Kim
                        NLP
                                55
                                       Delhi
                                             60000
                                                     10
In [67]:
          clean_data.info()
          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 6 entries, 0 to 5
          Data columns (total 6 columns):
                         Non-Null Count
               Column
                                           Dtype
           0
               Name
                         6 non-null
                                           object
               Domain
                         6 non-null
           1
                                           object
           2
               Age
                         6 non-null
                                           object
           3
               Location 6 non-null
                                           object
           4
               Salary
                          6 non-null
                                           object
           5
                          6 non-null
                                           object
               Exp
          dtypes: object(6)
          memory usage: 416.0+ bytes
          clean_data['Age']=clean_data['Age'].astype(int)
In [68]:
In [69]:
          clean_data.info()
          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 6 entries, 0 to 5
          Data columns (total 6 columns):
               Column
                         Non-Null Count Dtype
           #
          - - -
           0
               Name
                         6 non-null
                                           object
           1
               Domain
                         6 non-null
                                           object
           2
               Age
                         6 non-null
                                           int64
           3
               Location 6 non-null
                                           object
           4
               Salary
                          6 non-null
                                           object
           5
                          6 non-null
                                           object
               Exp
          dtypes: int64(1), object(5)
          memory usage: 416.0+ bytes
          clean_data['Salary'] = clean_data['Salary'].astype(int)
In [70]:
          clean_data['Exp'] = clean_data['Exp'].astype(int)
In [71]:
          clean_data.info()
          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 6 entries, 0 to 5
          Data columns (total 6 columns):
                         Non-Null Count Dtype
               Column
          - - -
           0
               Name
                         6 non-null
                                           object
           1
               Domain
                         6 non-null
                                           object
           2
                         6 non-null
                                           int64
               Age
           3
               Location 6 non-null
                                           object
           4
                         6 non-null
               Salary
                                           int64
           5
               Exp
                          6 non-null
                                           int64
```

clean_data

In [66]:

```
memory usage: 416.0+ bytes
          clean_data['Name'] = clean_data['Name'].astype('category')
In [72]:
          clean_data['Domain'] = clean_data['Domain'].astype('category')
          clean_data['Location'] = clean_data['Location'].astype('category')
In [73]:
          clean_data.info()
          <class 'pandas.core.frame.DataFrame'>
          RangeIndex: 6 entries, 0 to 5
          Data columns (total 6 columns):
                Column
                           Non-Null Count
           #
                                              Dtype
          - - -
           0
                Name
                           6 non-null
                                              category
           1
                Domain
                           6 non-null
                                              category
           2
                           6 non-null
                                              int64
                Age
           3
                Location 6 non-null
                                              category
                Salary
           4
                           6 non-null
                                              int64
           5
                Exp
                           6 non-null
                                              int64
          dtypes: category(3), int64(3)
          memory usage: 934.0 bytes
          clean_data
In [74]:
Out[74]:
             Name
                       Domain
                               Age
                                     Location
                                              Salary Exp
              Mike
                    Datascience
                                 34
                                      Mumbai
                                                5000
                                                        2
                                               10000
                                                        3
             Teddy
                        Testing
                                 45
                                    Bangalore
          2
              Umar
                    Dataanalyst
                                 50
                                               15000
                                                        4
                                    Bangalore
              Jane
                       Analytics
                                 50
                                     Hyderbad
                                               20000
                                                        4
             Uttam
                       Statistics
                                    Bangalore
                                               30000
                                                        5
                                 67
          5
               Kim
                          NLP
                                 55
                                         Delhi
                                               60000
                                                       10
          clean_data.to_csv('clean_data.csv')
In [75]:
          import os
In [76]:
          os.getcwd()
           '/content'
Out[76]:
          clean_data
In [77]:
                       Domain
                               Age
                                     Location
Out[77]:
             Name
                                              Salary
                                                     Exp
          0
                                                5000
                                                        2
              Mike
                    Datascience
                                 34
                                      Mumbai
                                                        3
             Teddy
                        Testing
                                 45
                                    Bangalore
                                               10000
          2
              Umar
                    Dataanalyst
                                 50
                                    Bangalore
                                               15000
                                                        4
          3
              Jane
                       Analytics
                                     Hyderbad
                                               20000
                                                        4
                                                        5
          4
             Uttam
                       Statistics
                                 67
                                    Bangalore
                                               30000
          5
                          NLP
                                 55
                                         Delhi
                                               60000
                                                       10
               Kim
```

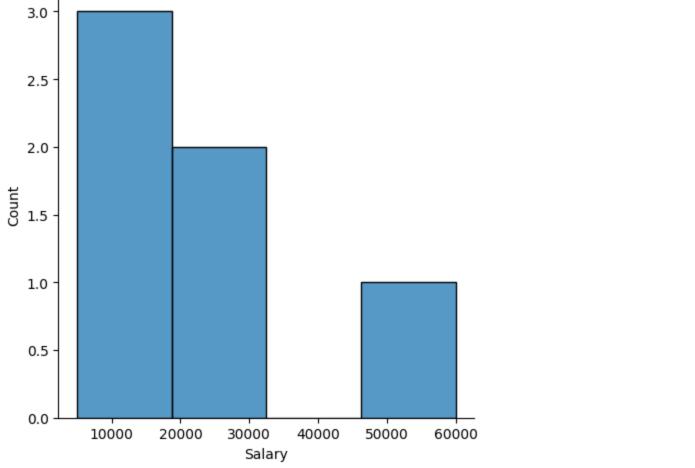
EDA TECHNIQUE APPLY **bold text**

dtypes: int64(3), object(3)

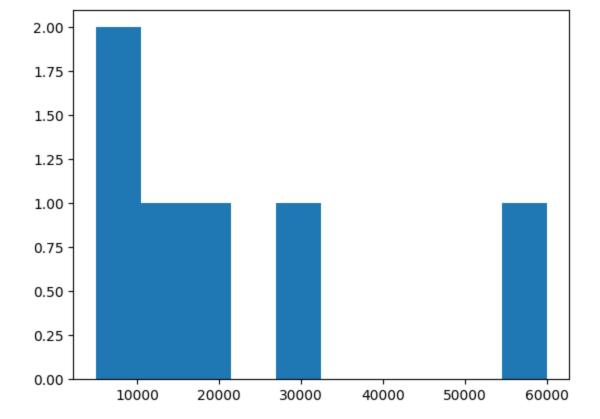
In [78]: **import** matplotlib.pyplot **as** plt

```
import warnings
In [79]:
          warnings.filterwarnings('ignore')
          clean_data['Salary']
In [80]:
Out[80]:
             Salary
              5000
             10000
          2
             15000
             20000
             30000
          5 60000
         dtype: int64
In [83]:
          vist1=sns.displot(clean_data['Salary'])
             3.0
             2.5
             2.0
```

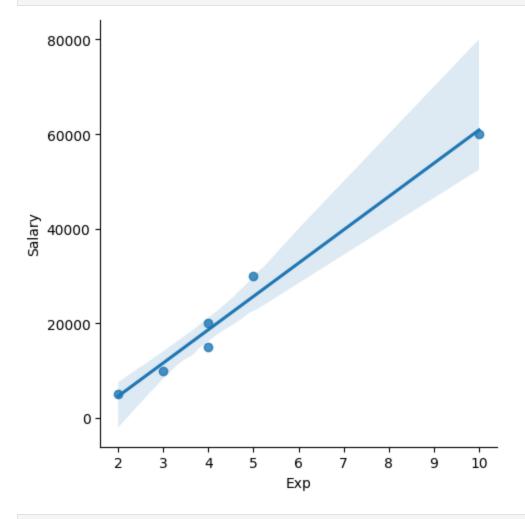
import seaborn as sns



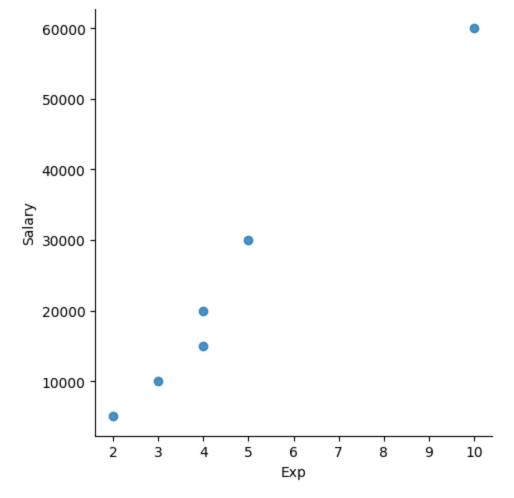
```
In [84]: vist2=plt.hist(clean_data['Salary'])
```



In [85]: vis3=sns.lmplot(data=clean_data, x='Exp', y='Salary')



In [87]: vis4=sns.lmplot(data=clean_data, x='Exp', y='Salary', fit_reg=False)



In [88]:	clean_data	a[:]				
Out[88]:	Name	Domain	Age	Location	Salary	Exp

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50	Bangalore	15000	4
3	Jane	Analytics	50	Hyderbad	20000	4
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [89]: clean_data[0:6:2]

Out[89]:		Name	Domain	Age	Location	Salary	Exp
	0 Mike		Datascience	34	Mumbai	5000	2
	2 Umar		Dataanalyst	50	Bangalore	15000	4
	4	Uttam	Statistics	67	Bangalore	30000	5

In [91]: clean_data[::-1]

Out[91]:		Name	Domain	Age	Location	Salary	Exp
	5	Kim	NLP	55	Delhi	60000	10
	4	Uttam	Statistics	67	Bangalore	30000	5
	3	Jane	Analytics	50	Hyderbad	20000	4

```
In [92]:
           clean_data.columns
           Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
Out[92]:
           X_iv=clean_data[['Name', 'Domain', 'Location', 'Age', 'Exp']]
In [93]:
                        Domain
                                  Location
                                           Age
                                                Exp
Out[93]:
              Name
                                                   2
               Mike
                     Datascience
                                  Mumbai
                                             34
              Teddy
                                 Bangalore
                                             45
                                                   3
                         Testing
           2
              Umar
                     Dataanalyst
                                 Bangalore
                                             50
                                                   4
               Jane
                        Analytics
                                 Hyderbad
                                             50
                                                   4
                                                   5
              Uttam
                        Statistics
                                 Bangalore
                                             67
                Kim
                           NLP
                                     Delhi
                                             55
                                                  10
           y_dv=clean_data['Salary']
In [94]:
           y_dv
Out[94]:
              Salary
               5000
              10000
              15000
              20000
              30000
              60000
          dtype: int64
In [95]:
Out[95]:
              Name
                        Domain
                                 Age
                                       Location Salary
                                                         Exp
                                                  5000
                                                           2
               Mike
                     Datascience
                                  34
                                        Mumbai
              Teddy
                         Testing
                                  45
                                       Bangalore
                                                 10000
                                                           3
              Umar
                     Dataanalyst
                                 NaN
                                           NaN
                                                 15000
                                                           4
                                 NaN
                                       Hyderbad
                                                 20000
                                                        NaN
               Jane
                        Analytics
                                                 30000
                                                           5
           4
              Uttam
                        Statistics
                                  67
                                           NaN
                           NLP
                                  55
                                           Delhi
                                                 60000
                                                          10
                Kim
In [96]:
           clean_data
```

Umar

Teddy

Out[96]:

Name

Domain Age

Location Salary

Mike

50

45

34

Dataanalyst

Datascience

Testing

Bangalore

Bangalore

Mumbai

15000

10000

5000

4

3

2

	Ū	WIIKC	Datascience	54	Widilibo	. 3000					
	1	Teddy	Testing	45	Bangalore	e 10000	3				
	2	Umar	Dataanalyst	50	Bangalore	e 15000	4				
	3	Jane	Analytics	50	Hyderbad	d 20000	4				
	4	Uttam	Statistics	67	Bangalore	e 30000	5				
	5	Kim	NLP	55	Delh	ni 60000	10				
In [97]:	Χ_	iv									
Out[97]:		Name	Domain	Loca	ation Age	е Ехр					
	0	Mike	Datascience	Mu	ımbai 34	4 2					
	1	Teddy	Testing	Bang	alore 4	5 3					
	2	Umar	Dataanalyst	Bang	alore 50	0 4					
	3	Jane	Analytics	Hyde	erbad 50	0 4					
	4	Uttam	Statistics	Bang	alore 6	7 5					
	5	Kim	NLP		Delhi 5	5 10					
In [98]:	у_	_dv									
Out[98]:		Salary									
	0	5000	_								
	1	10000									
	2	15000									
	3	20000									
	4	30000									
	5	60000									
	dtv	pe: inte	34								
	_										
In [99]:	cl	.ean_da	ata								
Out[99]:		Name	Domain	Age	Location	n Salary	Exp				
	0	Mike	Datascience	34	Mumba		2				
	1	Teddy	Testing	45	Bangalore		3				
	2	Umar	Dataanalyst	50	Bangalore		4				
	3	Jane	Analytics	50	Hyderbad	d 20000	4				
	4	Uttam	Statistics	67	Bangalore	e 30000	5				
	5	Kim	NLP	55	Delh	ni 60000	10				
In [108		ıputati ıputati	ion=pd.get _. ion	_dumm	ies(clea	an_data)					
Out[108]:		Age	Salary Exp	Nam	e_Jane I	Name_Kim	Name_Mike	Name_Teddy	Name_Umar	Name_Uttam	Domain_

Mike Datascience 34 Mumbai 5000 2

1	45	10000	3	False	False	False	True	False	False	
2	50	15000	4	False	False	False	False	True	False	
3	50	20000	4	True	False	False	False	False	False	
4	67	30000	5	False	False	False	False	False	True	
5	55	60000	10	False	True	False	False	False	False	

In [101... clean_data

Out[101]:

	Name	Domain	Age	Location	Salary	Exp
0	Mike	Datascience	34	Mumbai	5000	2
1	Teddy	Testing	45	Bangalore	10000	3
2	Umar	Dataanalyst	50	Bangalore	15000	4
3	Jane	Analytics	50	Hyderbad	20000	4
4	Uttam	Statistics	67	Bangalore	30000	5
5	Kim	NLP	55	Delhi	60000	10

In [102... imputation

Out[102]:

	Age	Salary	Exp	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar	Name_Uttam	Domain_/
0	34	5000	2	False	False	True	False	False	False	
1	45	10000	3	False	False	False	True	False	False	
2	50	15000	4	False	False	False	False	True	False	
3	50	20000	4	True	False	False	False	False	False	
4	67	30000	5	False	False	False	False	False	True	
5	55	60000	10	False	True	False	False	False	False	

Out[103]:

Name	Domain	Age	Location	Salary	Exp
Mike	Datascience	34	Mumbai	5000	2
Teddy	Testing	45	Bangalore	10000	3
Umar	Dataanalyst	50	Bangalore	15000	4
Jane	Analytics	50	Hyderbad	20000	4
Uttam	Statistics	67	Bangalore	30000	5
Kim	NLP	55	Delhi	60000	10
	Mike Teddy Umar Jane Uttam	Mike Datascience Teddy Testing Umar Dataanalyst Jane Analytics Uttam Statistics	Mike Datascience 34 Teddy Testing 45 Umar Dataanalyst 50 Jane Analytics 50 Uttam Statistics 67	Mike Datascience 34 Mumbai Teddy Testing 45 Bangalore Umar Dataanalyst 50 Bangalore Jane Analytics 50 Hyderbad Uttam Statistics 67 Bangalore	MikeDatascience34Mumbai5000TeddyTesting45Bangalore10000UmarDataanalyst50Bangalore15000JaneAnalytics50Hyderbad20000UttamStatistics67Bangalore30000

In [104... imputation

Out[104]

:		Age	Salary	Exp	Name_Jane	Name_Kim	Name_Mike	Name_Teddy	Name_Umar	Name_Uttam	Domain_/
	0	34	5000	2	False	False	True	False	False	False	
	1	45	10000	3	False	False	False	True	False	False	
	2	50	15000	4	False	False	False	False	True	False	

```
20000
                                            False
                                                          False
                                                                         False
                                                                                        False
                                                                                                       False
3
    50
                                True
          30000
                               False
                                            False
                                                          False
                                                                         False
                                                                                        False
                                                                                                        True
5
    55 60000
                   10
                               False
                                             True
                                                          False
                                                                         False
                                                                                        False
                                                                                                       False
```

```
clean_data=clean_data.astype(int)
In [105...
         clean_data
         ValueError
                                                    Traceback (most recent call last)
         /usr/local/lib/python3.10/dist-packages/pandas/core/arrays/categorical.py in astype(sel
         f, dtype, copy)
             563
                             try:
         --> 564
                                 new_cats = new_cats.astype(dtype=dtype, copy=copy)
             565
                                 fill_value = self.categories._na_value
         ValueError: invalid literal for int() with base 10: 'Analytics'
         During handling of the above exception, another exception occurred:
         ValueError
                                                    Traceback (most recent call last)
         <ipython-input-105-6d36147c8c60> in <cell line: 1>()
         ----> 1 clean_data=clean_data.astype(int)
               2 clean_data
         /usr/local/lib/python3.10/dist-packages/pandas/core/generic.py in astype(self, dtype, co
         py, errors)
            6532
                         else:
            6533
                             # else, only a single dtype is given
         -> 6534
                             new_data = self._mgr.astype(dtype=dtype, copy=copy, errors=errors)
            6535
                             res = self._constructor_from_mgr(new_data, axes=new_data.axes)
            6536
                             return res.__finalize__(self, method="astype")
         /usr/local/lib/python3.10/dist-packages/pandas/core/internals/managers.py in astype(sel
         f, dtype, copy, errors)
             412
                             copy = False
             413
         --> 414
                         return self.apply(
             415
                             "astype",
             416
                             dtype=dtype,
         /usr/local/lib/python3.10/dist-packages/pandas/core/internals/managers.py in apply(self,
          f, align_keys, **kwargs)
             352
                                  applied = b.apply(f, **kwargs)
             353
                             else:
         --> 354
                                  applied = getattr(b, f)(**kwargs)
             355
                             result_blocks = extend_blocks(applied, result_blocks)
             356
         /usr/local/lib/python3.10/dist-packages/pandas/core/internals/blocks.py in astype(self,
          dtype, copy, errors, using_cow)
             614
                         values = self.values
             615
         --> 616
                         new_values = astype_array_safe(values, dtype, copy=copy, errors=errors)
             617
             618
                         new_values = maybe_coerce_values(new_values)
         /usr/local/lib/python3.10/dist-packages/pandas/core/dtypes/astype.py in astype_array_saf
         e(values, dtype, copy, errors)
             236
             237
                     try:
         --> 238
                         new_values = astype_array(values, dtype, copy=copy)
```

240

except (ValueError, TypeError):

e.g. _astype_nansafe can fail on object-dtype of strings

```
/usr/local/lib/python3.10/dist-packages/pandas/core/dtypes/astype.py in astype_array(val
ues, dtype, copy)
           if not isinstance(values, np.ndarray):
    178
    179
               # i.e. ExtensionArray
--> 180
               values = values.astype(dtype, copy=copy)
   181
    182
          else:
/usr/local/lib/python3.10/dist-packages/pandas/core/arrays/categorical.py in astype(sel
f, dtype, copy)
    573
                    ):
    574
                        msg = f"Cannot cast {self.categories.dtype} dtype to {dtype}"
--> 575
                        raise ValueError(msg)
    576
                    result = take_nd(
    577
ValueError: Cannot cast object dtype to int64
```

In []: