```
import pandas as pd
In [29]:
In [30]:
         pd.__version__
Out[30]: '2.2.3'
         emp = pd.read_excel(r"C:\Users\chamb\Downloads\Rawdata.xlsx")
          emp
Out[31]:
              Name
                           Domain
                                       Age
                                              Location
                                                          Salary
                                                                    Exp
          0
               Mike
                      Datascience#$ 34 years
                                                         5^00#0
                                              Mumbai
                                                                     2+
            Teddy^
          1
                            Testing
                                      45' yr
                                             Bangalore
                                                       10%%000
                                                                     <3
          2
             Uma#r Dataanalyst^^#
                                       NaN
                                                 NaN
                                                        1$5%000
                                                                  4> yrs
                        Ana^^lytics
                                            Hyderbad
          3
               Jane
                                       NaN
                                                         2000^0
                                                                    NaN
             Uttam*
          4
                           Statistics
                                      67-yr
                                                 NaN
                                                         30000-
                                                                 5+ year
          5
                Kim
                               NLP
                                       55yr
                                                 Delhi
                                                        6000^$0
                                                                    10+
In [32]:
         id(emp)
Out[32]:
          1711953997136
In [33]:
         emp.shape
Out[33]: (6, 6)
In [34]:
          emp.columns
Out[34]: Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
In [35]:
          emp.head
Out[35]:
          <bound method NDFrame.head of</pre>
                                                                                Location
                                              Name
                                                             Domain
                                                                          Age
                      Exp
          Salary
                                                                          2+
          0
               Mike
                      Datascience#$ 34 years
                                                   Mumbai
                                                             5^00#0
            Teddy^
                                     45' yr Bangalore
                                                           10%%000
          1
                            Testing
                                                                          <3
          2
             Uma#r
                     Dataanalyst^^#
                                           NaN
                                                      NaN
                                                           1$5%000
                                                                      4> yrs
                        Ana^^lytics
                                           NaN
          3
               Jane
                                                 Hyderbad
                                                             2000^0
                                                                         NaN
            Uttam*
                         Statistics
                                         67-yr
                                                      NaN
                                                             30000-
                                                                     5+ year
          5
                Kim
                                NLP
                                                    Delhi
                                                            6000^$0
                                                                         10+>
                                          55yr
In [36]: emp.tail
Out[36]: <bound method NDFrame.tail of
                                                             Domain
                                              Name
                                                                                Location
                                                                          Age
          Salary
                      Exp
               Mike
                      Datascience#$ 34 years
                                                             5^00#0
                                                                          2+
          0
                                                   Mumbai
          1
            Teddv^
                            Testing 45' yr Bangalore
                                                           10%%000
                                                                          <3
          2
              Uma#r
                     Dataanalyst^^#
                                           NaN
                                                           1$5%000
                                                      NaN
                                                                      4> yrs
                        Ana^^lytics
          3
               Jane
                                           NaN
                                                 Hyderbad
                                                             2000^0
                                                                         NaN
                         Statistics
          4
             Uttam*
                                         67-yr
                                                      NaN
                                                             30000-
                                                                     5+ year
          5
                Kim
                                NLP
                                          55yr
                                                    Delhi
                                                           6000^$0
                                                                         10+>
```

```
In [37]: 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
                                          object
         1
              Domain 6 non-null
         2
              Age
                       4 non-null
                                          object
         3
              Location 4 non-null
                                          object
         4
              Salary
                        6 non-null
                                          object
         5
                         5 non-null
                                          object
              Exp
         dtypes: object(6)
        memory usage: 420.0+ bytes
In [38]: emp.isnull()
Out[38]:
             Name Domain Age Location Salary
                                                       Exp
          0
              False
                        False
                              False
                                        False
                                                False
                                                      False
               False
          1
                        False False
                                        False
                                                False False
          2
              False
                        False True
                                        True
                                                False False
          3
              False
                        False True
                                        False
                                                False
                                                      True
          4
              False
                        False False
                                        True
                                                False False
          5
               False
                        False
                              False
                                        False
                                                False False
In [39]:
          emp.isna()
Out[39]:
                     Domain
             Name
                              Age Location
                                              Salary
                                                       Exp
          0
              False
                              False
                                        False
                        False
                                                False
                                                      False
               False
                                        False
          1
                        False
                              False
                                                False
                                                      False
          2
              False
                        False
                                        True
                                                False
                                                      False
                              True
              False
                                        False
          3
                        False
                              True
                                                False
                                                      True
          4
              False
                        False
                              False
                                        True
                                                False
                                                      False
          5
               False
                                        False
                        False
                              False
                                                False False
          emp.isnull().sum()
In [40]:
Out[40]:
          Name
                       0
          Domain
                       0
          Age
                       2
          Location
                       2
                       0
          Salary
          Exp
          dtype: int64
          emp['Name']
In [41]:
```

```
Out[41]: 0 Mike
1 Teddy^
2 Uma#r
3 Jane
4 Uttam*
5 Kim
Name: Name, dtype: object
```

Data Cleaning

```
In [42]: emp
Out[42]:
              Name
                                              Location
                                                          Salary
                           Domain
                                       Age
                                                                     Exp
                      Datascience#$ 34 years
                                                                     2+
          0
               Mike
                                               Mumbai
                                                         5^00#0
            Teddy^
                                      45' yr Bangalore
                            Testing
                                                       10%%000
                                                                      <3
             Uma#r Dataanalyst^^#
          2
                                       NaN
                                                  NaN
                                                        1$5%000
                                                                   4> yrs
                        Ana^^lytics
                                       NaN Hyderbad
          3
                                                         2000^0
                                                                    NaN
               Jane
             Uttam*
                           Statistics
                                       67-yr
                                                          30000- 5+ year
                                                  NaN
          5
                Kim
                               NLP
                                       55yr
                                                 Delhi
                                                        6000^$0
                                                                     10+
In [43]:
         emp['Name']
Out[43]: 0
                 Mike
          1
               Teddy^
          2
               Uma#r
          3
                 Jane
               Uttam*
          4
                  Kim
          Name: Name, dtype: object
         emp['Name'] = emp['Name'].str.replace(r'\W', '', regex=True) #\W= non word chara
In [45]: emp['Name']
Out[45]: 0
                Mike
          1
               Teddy
          2
                Umar
          3
                Jane
          4
               Uttam
                 Kim
          Name: Name, dtype: object
In [46]:
         emp
```

```
Out[46]:
             Name
                          Domain
                                       Age
                                             Location
                                                         Salary
                                                                   Exp
          0
              Mike
                     Datascience#$ 34 years
                                             Mumbai
                                                        5^00#0
                                                                    2+
             Teddy
                           Testing
                                     45' yr Bangalore 10%%000
                                                                     <3
          2
             Umar
                   Dataanalyst^^#
                                      NaN
                                                 NaN
                                                       1$5%000
                                                                 4> yrs
          3
              Jane
                       Ana^^lytics
                                      NaN
                                            Hyderbad
                                                        2000^0
                                                                   NaN
            Uttam
                          Statistics
                                      67-yr
                                                 NaN
                                                         30000- 5+ year
               Kim
                              NLP
                                       55yr
                                                Delhi
                                                       6000^$0
                                                                    10+
         emp['Domain']
In [47]:
Out[47]: 0
                Datascience#$
                      Testing
          2
               Dataanalyst^^#
          3
                  Ana^^lytics
                   Statistics
          4
                          NLP
          Name: Domain, dtype: object
         emp['Domain'] = emp['Domain'].str.replace(r'\W', '', regex=True)
In [48]:
In [49]:
         emp['Domain']
          0
Out[49]:
               Datascience
          1
                   Testing
          2
               Dataanalyst
          3
                 Analytics
          4
                Statistics
                       NLP
          Name: Domain, dtype: object
In [50]:
         emp['Age'] = emp['Age'].str.replace(r'\W', '', regex=True)
In [51]:
         emp['Age']
Out[51]: 0
               34years
          1
                  45yr
          2
                   NaN
          3
                   NaN
          4
                  67yr
          5
                  55yr
          Name: Age, dtype: object
         emp['Age'] = emp['Age'].str.extract('(\\d+)') #r(r'(\\d+)' for extracting text
In [52]:
In [53]:
         emp['Age']
Out[53]:
          0
                34
                45
          1
          2
               NaN
          3
               NaN
          4
                67
                55
          Name: Age, dtype: object
```

```
In [54]: emp['Location'] = emp['Location'].str.replace(r'\W', '', regex=True)
In [55]: emp['Location']
Out[55]: 0
                  Mumbai
          1
               Bangalore
          2
                     NaN
               Hyderbad
          3
          4
                     NaN
          5
                   Delhi
          Name: Location, dtype: object
In [56]:
         emp
Out[56]:
             Name
                                      Location
                       Domain
                                Age
                                                  Salary
                                                             Exp
          0
              Mike Datascience
                                 34
                                                 5^00#0
                                                              2+
                                       Mumbai
                                               10%%000
             Teddy
                                 45
                                                              <3
                                     Bangalore
                        Testing
                                                1$5%000
          2
             Umar
                    Dataanalyst NaN
                                          NaN
                                                           4> yrs
                                     Hyderbad
                                                 2000^0
          3
                      Analytics
                                NaN
                                                            NaN
              Jane
             Uttam
                       Statistics
                                 67
                                          NaN
                                                  30000-
                                                          5+ year
          5
                          NLP
                                                6000^$0
               Kim
                                 55
                                         Delhi
                                                             10+
In [57]: emp['Salary'] = emp['Salary'].str.replace(r'\W', '', regex=True)
In [58]:
         emp['Salary']
Out[58]:
                5000
          1
               10000
          2
               15000
          3
               20000
          4
               30000
          5
               60000
          Name: Salary, dtype: object
In [59]: emp['Exp'] = emp['Exp'].str.extract('(\\d+)')
In [60]:
         emp['Exp']
Out[60]:
          0
                 2
          1
                 3
          2
                 4
          3
               NaN
                 5
          4
          5
                10
          Name: Exp, dtype: object
In [61]: emp
```

Out[61]:		Name	Domain	Age	Location	Salary	Ехр
	0	Mike	Datascience	34	Mumbai	5000	2
	1	Teddy	Testing	45	Bangalore	10000	3
	2	Umar	Dataanalyst	NaN	NaN	15000	4
	3	Jane	Analytics	NaN	Hyderbad	20000	NaN
	4	Uttam	Statistics	67	NaN	30000	5
	5	Kim	NLP	55	Delhi	60000	10
In [62]:	cl	ean_dat	a = emp.copy	y()			
In [63]:	cl	ean_dat	a				
<pre>In [63]: Out[63]:</pre>	cl	ean_dat Name	a Domain	Age	Location	Salary	Ехр
	c1 0	Name		Age 34	Location Mumbai	Salary 5000	Exp 2
		Name	Domain				
	0	Name Mike	Domain Datascience	34 45	Mumbai	5000	2
	0	Name Mike Teddy	Domain Datascience Testing Dataanalyst	34 45 NaN	Mumbai Bangalore	5000 10000 15000	2
	0 1 2	Name Mike Teddy Umar	Domain Datascience Testing Dataanalyst	34 45 NaN	Mumbai Bangalore NaN	5000 10000 15000	2 3 4

eda techniques

In [64]:	clean_data						
Out[64]:		Name	Domain	Age	Location	Salary	Ехр
	0	Mike	Datascience	34	Mumbai	5000	2
	1	Teddy	Testing	45	Bangalore	10000	3
	2	Umar	Dataanalyst	NaN	NaN	15000	4
	3	Jane	Analytics	NaN	Hyderbad	20000	NaN
	4	Uttam	Statistics	67	NaN	30000	5
	5	Kim	NLP	55	Delhi	60000	10
In [65]:	<pre>clean_data.isnull().sum()</pre>						

```
Out[65]: Name
                                                                                                                                            0
                                                               Domain
                                                                                                                                            0
                                                                                                                                            2
                                                               Age
                                                                                                                                           2
                                                               Location
                                                               Salary
                                                               Exp
                                                                                                                                            1
                                                               dtype: int64
In [66]: clean_data['Age']
Out[66]: 0
                                                                                                      34
                                                                                                     45
                                                               1
                                                               2
                                                                                               NaN
                                                               3
                                                                                             NaN
                                                               4
                                                                                                     67
                                                                                                      55
                                                               5
                                                               Name: Age, dtype: object
In [67]: import numpy as np
                                                        clean_data['Age'] = clean_data['Age'].fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.mean(pd.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age']).fillna(np.to_numeric(clean_data['Age'])).fillna(np.to_numeric(clean_data['Age'])).fillna(np.to_numeric(clean_data['Age'])).fillna(np.to_numeric(clean_data['Age'])).fillna(np.to_numeric(clean_data['Age'])
In [68]:
In [69]:
                                                             clean_data['Age']
                                                               0
                                                                                                                  34
Out[69]:
                                                                                                                  45
                                                               1
                                                               2
                                                                                              50.25
                                                               3
                                                                                               50.25
                                                               4
                                                                                                                  67
                                                               5
                                                                                                                  55
                                                               Name: Age, dtype: object
In [70]: clean_data['Exp'] = clean_data['Exp'].fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.mean(pd.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp']).fillna(np.to_numeric(clean_data['Exp'])).fillna(np
In [71]: clean_data['Exp']
Out[71]:
                                                              0
                                                                                                            2
                                                               1
                                                                                                            3
                                                               2
                                                                                                            4
                                                                                               4.8
                                                               3
                                                               4
                                                                                                            5
                                                               5
                                                                                                      10
                                                               Name: Exp, dtype: object
In [72]:
                                                        clean_data['Location'].isnull().sum()
Out[72]: np.int64(2)
In [73]:
                                                             clean_data['Location']
                                                                                                                 Mumbai
Out[73]: 0
                                                               1
                                                                                               Bangalore
                                                               2
                                                                                                                                     NaN
                                                               3
                                                                                                    Hyderbad
                                                               4
                                                                                                                                     NaN
                                                                                                                         Delhi
                                                               Name: Location, dtype: object
```

```
clean_data['Location'] = clean_data['Location'].fillna(clean_data['Location'].mod
In [75]: clean_data['Location']
Out[75]: 0
                Mumbai
         1
              Bangalore
         2
              Bangalore
         3
             Hyderbad
         4
              Bangalore
         5
                 Delhi
         Name: Location, dtype: object
In [76]: 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 object
3 Location 6 non-null object
4 Salary 6 non-null object
        5
            Exp
                      6 non-null
                                     object
       dtypes: object(6)
       memory usage: 420.0+ bytes
In [77]: clean_data['Age'] = clean_data['Age'].astype(int)
In [78]: 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
                   6 non-null
        1
            Domain
                                     object
                 6 non-null
        2 Age
                                     int64
        3 Location 6 non-null
                                     object
        4 Salary
                     6 non-null
                                     object
                      6 non-null
                                     object
            Exp
       dtypes: int64(1), object(5)
       memory usage: 420.0+ bytes
         clean_data['Salary'] = clean_data['Salary'].astype(int)
In [80]: 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
                                   int64
        5
                   6 non-null
                                   object
           Exp
       dtypes: int64(2), object(4)
       memory usage: 420.0+ bytes
In [81]: clean_data['Exp'] = clean_data['Exp'].astype(int)
In [82]: 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
                   6 non-null
                                 int64
        2 Age
        3 Location 6 non-null
                                   object
        4 Salary 6 non-null
                                   int64
                   6 non-null
        5
           Exp
                                   int64
       dtypes: int64(3), object(3)
       memory usage: 420.0+ bytes
        clean_data['Name'] = clean_data['Name'].astype('category')
In [83]:
        clean_data['Domain'] = clean_data['Domain'].astype('category')
        clean_data['Location'] = clean_data['Location'].astype('category')
In [84]: 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
          Domain 6 non-null
                                 category
        1
        2 Age 6 non-null
                                   int64
        3 Location 6 non-null
                                   category
        4
                    6 non-null
           Salary
                                   int64
        5
           Exp
                    6 non-null
                                   int64
       dtypes: category(3), int64(3)
       memory usage: 938.0 bytes
In [85]:
       clean_data
```

Out[85]:		Name	Domain	Age	Location	Salary	Ехр	
	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 [86]:	<pre>clean_data.to_csv('clean_data.csv')</pre>							
In [87]:	<pre>import os os.getcwd() # to get path of the file</pre>							

Out[87]: 'C:\\Users\\chamb'