In [7]:
```
import pandas as pd
pd.__version__
```

Out[7]:
```
'2.2.2'
```

In [ ]:
```
#pip install --upgrade openpyxl
```

In [2]:
```
df = pd.read_excel(r'C:\Users\ADMIN\Downloads\Rawdata.xlsx')
```

In [3]:
```
df
```

Out[3]:

| | 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 [8]:
```
id(df)
```

Out[8]:
```
1760629027184
```

In [9]:
```
df.columns
```

Out[9]:
```
Index(['Name', 'Domain', 'Age', 'Location', 'Salary', 'Exp'], dtype='object')
```

In [11]:
```
df.shape
```

Out[11]:
```
(6, 6)
```

In [12]:
```
df.head()
```

Out[12]:

| | 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 |

In [13]:
```
df.tail()
```

Out[13]:

| | Name | Domain | Age | Location | Salary | Exp |
|---|---|---|---|---|---|---|
| **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 [14]:
```
df.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       4 non-null      object
 3   Location  4 non-null      object
 4   Salary    6 non-null      object
 5   Exp       5 non-null      object
dtypes: object(6)
memory usage: 420.0+ bytes
```

In [15]:
```
df.isnull()
```

Out[15]:

| | Name | Domain | Age | Location | Salary | Exp |
|---|---|---|---|---|---|---|
| **0** | False | False | False | False | False | False |
| **1** | False | 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 [16]:
```
df.isna()
```

Out[16]:

| | Name | Domain | Age | Location | Salary | Exp |
|---|---|---|---|---|---|---|
| **0** | False | False | False | False | False | False |
| **1** | False | 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 [18]:
```
df.isnull().sum()  # gives the count of null values
```

Out[18]:
```
Name        0
Domain      0
Age         2
Location    2
Salary      0
Exp         1
dtype: int64
```

# Data Cleaning

In [19]:
```
df
```

Out[19]:

| | 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 [20]:
```
df['Name']
```

Out[20]:
```
0      Mike
1    Teddy^
2     Uma#r
3      Jane
4    Uttam*
5       Kim
Name: Name, dtype: object
```

In [23]:
```
df['Name'] = df['Name'].str.replace(r'\W','',regex=True)   # nonword character
```

In [22]:
```
df['Name']
```

Out[22]:
```
0     Mike
1    Teddy
2     Umar
3     Jane
4    Uttam
5      Kim
Name: Name, dtype: object
```

In [24]:
```
df['Domain']
```

Out[24]:
```
0     Datascience#$
1           Testing
2    Dataanalyst^^#
3       Ana^^lytics
4        Statistics
5               NLP
Name: Domain, dtype: object
```

In [25]:
```
df['Domain'] = df['Domain'].str.replace(r'\W','',regex=True)    # nonword character
```

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

Out[26]:
```
0    Datascience
1        Testing
2    Dataanalyst
3      Analytics
4     Statistics
5            NLP
Name: Domain, dtype: object
```

In [28]:
```
df['Location']
```

Out[28]:
```
0       Mumbai
1    Bangalore
2          NaN
3     Hyderbad
4          NaN
5        Delhi
Name: Location, dtype: object
```

In [27]:
```
df['Age']
```

Out[27]:
```
0    34 years
1      45' yr
2         NaN
3         NaN
4       67-yr
5        55yr
Name: Age, dtype: object
```

In [29]:
```
df['Age'] =df['Age'].str.replace(r'\W','',regex=True)
```

In [30]:
```
df['Age']
```

Out[30]:
```
0    34years
1       45yr
2        NaN
3        NaN
4       67yr
5       55yr
Name: Age, dtype: object
```

In [33]:
```
df['Age'] =df['Age'].str.extract('(\\d+)')   # \\d is used to exract the string
```

In [32]:
```
df['Age']
```

Out[32]:
```
0     34
1     45
2    NaN
3    NaN
4     67
5     55
Name: Age, dtype: object
```

In [34]:
```
df
```

Out[34]:

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

In [35]:
```
df['Salary'] = df['Salary'].str.replace(r'\W','',regex=True)
```

In [36]:
```
df['Exp'] =df['Exp'].str.extract('(\\d+)')
```

In [37]:
```
df
```

Out[37]:

| | Name | Domain | Age | Location | Salary | Exp |
|---|---|---|---|---|---|---|
| **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 [39]:
```
clean_data = df.copy()
```

In [40]:
```
clean_data
```

Out[40]:

| | Name | Domain | Age | Location | Salary | Exp |
|---|---|---|---|---|---|---|
| **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 |