In [38]:

```
import pandas as pd
import numpy as np
```

In [39]:

```
data = pd.read_excel(″D:\KSR\Python\ExcelFile\Car_Data_mv.xlsx″ )
```

In [40]:

```
data.head(5)
```

Out[40]:

| | Car_Name | Year | Selling_Price | Present_Price | Kms_Driven | Fuel_Type | Seller_Type |
|---|---|---|---|---|---|---|---|
| 0 | ritz | 2014 | 3.35 | 5.59 | 27000 | Petrol | Dealer |
| 1 | sx4 | 2013 | 4.75 | 9.54 | 43000 | Diesel | Dealer |
| 2 | ciaz | 2017 | 7.25 | 9.85 | 6900 | Petrol | Dealer |
| 3 | wagon r | 2011 | 2.85 | 4.15 | 5200 | Petrol | Dealer |
| 4 | swift | 2014 | 4.6 | 6.87 | 42450 | Diesel | Dealer |

In [41]:

```
data.shape
```

Out[41]:

```
(301, 9)
```

In [42]:

```
list(data.columns)
```

Out[42]:

```
['Car_Name',
 'Year',
 'Selling_Price',
 'Present_Price',
 'Kms_Driven',
 'Fuel_Type',
 'Seller_Type',
 'Transmission',
 'Car_Owner']
```

In [43]:

```
#to find the missing value
```

In [44]:

```
data.isna().sum()
```

Out[44]:

```
Car_Name          6
Year              7
Selling_Price     0
Present_Price     0
Kms_Driven       15
Fuel_Type         6
Seller_Type      13
Transmission      8
Car_Owner         5
dtype: int64
```

In  [45]:

```
#missing value percentage
data.isna().sum() / data.shape[0] * 100
```

Out[45]:

```
Car_Name         1.993355
Year             2.325581
Selling_Price    0.000000
Present_Price    0.000000
Kms_Driven       4.983389
Fuel_Type        1.993355
Seller_Type      4.318937
Transmission     2.657807
Car_Owner        1.661130
dtype: float64
```

In  [100]:

```
#impute missing values(filling null values)
#if missing value % is > 75%(drop the column)
#if missing value % is > 30%(impute it )---> (fill it)
#if missing value % is btw (30 to 75) (check with the clients,
#reshare the data, can u send me the corrected data) suggestion?
```

In  [132]:

```
data = {
'Name': ["Kiran","Anand","Vinay","Rahul","Govind","Eashwar"],
'Age' : [25,None,30,None,26,28],
'Salary': [5000,3500,None,4500,5500,80000],
'Dept': ["HR","IT", np.nan, np.nan,"IT","HR"]
}
```

In [133]:

```
type(data)
```

Out[133]:

```
dict
```

In [134]:

```
data
```

Out[134]:

```
{'Name’: ['Kiran’, 'Anand’, 'Vinay’, 'Rahul’, 'Govind’, 'Eashwar’],
 'Age’: [25, None, 30, None, 26, 28],
 'Salary’: [5000, 3500, None, 4500, 5500, 80000],
 'Dept’: ['HR’, 'IT’, nan, nan, 'IT’, 'HR’]}
```

In [135]:

```
df = pd.DataFrame(data)
```

In [136]:

```
df
```

Out[136]:

| | Name | Age | Salary | Dept |
|---|---|---|---|---|
| **0** | Kiran | 25 | 5000 | HR |
| **1** | Anand | NaN | 3500 | IT |
| **2** | Vinay | 30 | NaN | NaN |
| **3** | Rahul | NaN | 4500 | NaN |
| **4** | Govind | 26 | 5500 | IT |
| **5** | Eashwar | 28 | 80000 | HR |

In [137]:

```
df.isna().sum() /df.shape[0] *100
```

Out[137]:

```
Name       0.000000
Age       33.333333
Salary    16.666667
Dept      33.333333
dtype: float64
```

In [138]:

```
df['Age'].mean()
```

Out[138]:

```
np.float64(27.25)
```

In [151]:

```
df['Age'] = df['Age'].fillna(df['Age'].mean())
```

In [152]:

```
df.head(6)
```

Out[152]:

| | Name | Age | Salary | Dept |
|---|---|---|---|---|
| 0 | Kiran | 25 | 5000 | HR |
| 1 | Anand | 27.25 | 3500 | IT |
| 2 | Vinay | 30 | 5000 | HR |
| 3 | Rahul | 27.25 | 4500 | HR |
| 4 | Govind | 26 | 5500 | IT |
| 5 | Eashwar | 28 | 80000 | HR |

In [153]:

```
df['Salary'].median()
```

Out[153]:

```
  5000
```

In [154]:

```
#mean/median/mode
```

In [155]:

```
df['Salary'] = df['Salary'].fillna(df['Salary'].median())
```

In [156]:

```
df
```

Out[156]:

| | Name | Age | Salary | Dept |
|---|---|---|---|---|
| 0 | Kiran | 25 | 5000 | HR |
| 1 | Anand | 27.25 | 3500 | IT |
| 2 | Vinay | 30 | 5000 | HR |

| | | | | |
|---|---|---|---|---|
| 3 | Rahul | 27.25 | 4500 | HR |
| 4 | Govind | 26 | 5500 | IT |
| 5 | Eashwar | 28 | 80000 | HR |

In [157]:

```
df.isna().sum()
```

Out[157]:

```
Name      0
Age       0
Salary    0
Dept      0
dtype: int64
```

In [158]:

```
#mode ---> most repeated times(nums/str)
```

In [159]:

```
df['Dept'].mode()
```

Out[159]:

```
0    HR
Name: Dept, dtype: object
```

In [164]:

```
df['Dept'] = df['Dept'].fillna(df['Dept'].mode()[0])
```

In [165]:

```
df
```

Out[165]:

| | Name | Age | Salary | Dept |
|---|---|---|---|---|
| 0 | Kiran | 25 | 5000 | HR |
| 1 | Anand | 27.25 | 3500 | IT |
| 2 | Vinay | 30 | 5000 | HR |
| 3 | Rahul | 27.25 | 4500 | HR |
| 4 | Govind | 26 | 5500 | IT |
| 5 | Eashwar | 28 | 80000 | HR |

In [ ]:

| Transmission | Car_Owner |
|---|---|
| Manual | 1st Owner |
| Manual | 1st Owner |
| Manual | 1st Owner |
| Manual | 1st Owner |
| Manual | 2nd Owner |