In [1]:
```
import numpy as np
import pandas as pd
```

In [2]:
```
data = {
    'Person': ['A', 'B', 'C'],
    'Age': [30, 50, None],
    'Sal': [20000, 80000, 30000]
}
```

In [3]:
```
df = pd.DataFrame(data)
```

In [4]:
```
df
```

Out[4]:

	Person	Age	Sal
0	A	30.0	20000
1	B	50.0	80000
2	C	NaN	30000

In [5]:
```
df['Age'] = df['Age'].fillna(df['Age'].mean())
```

In [6]:
```
df
```

Out[6]:

	Person	Age	Sal
0	A	30.0	20000
1	B	50.0	80000
2	C	40.0	30000

In [7]:
```
minAge = df['Age'].min()
maxAge = df['Age'].max()

ageScalingFactor = maxAge - minAge

df['Age'] = (df['Age'] - minAge) / ageScalingFactor
```

In [8]:
```
df
```

Out[8]:

	Person	Age	Sal
0	A	0.0	20000
1	B	1.0	80000
2	C	0.5	30000

In [9]:
```
minSal = df['Sal'].min()
maxSal = df['Sal'].max()
```

```
salScalingFactor = maxSal - minSal

df['Sal'] = round((df['Sal'] - minSal) / salScalingFactor, 1)
```

In [10]:
```
df
```

Out[10]:

	Person	Age	Sal
0	A	0.0	0.0
1	B	1.0	1.0
2	C	0.5	0.2

In [11]:
```
df2 = pd.read_csv(r'/content/ev.csv')
```

In [12]:
```
df2.head()
```

Out[12]:

	Series_reference	Period	Data_value	Suppressed	STATUS	UNITS	Magnit
0	ECTA.S19A1	2001.03	2462.5	NaN	F	Dollars	
1	ECTA.S19A1	2002.03	17177.2	NaN	F	Dollars	
2	ECTA.S19A1	2003.03	22530.5	NaN	F	Dollars	
3	ECTA.S19A1	2004.03	28005.1	NaN	F	Dollars	
4	ECTA.S19A1	2005.03	30629.6	NaN	F	Dollars	

In [13]:
```
df2.columns
```

Out[13]:
```
Index(['Series_reference', 'Period', 'Data_value', 'Suppressed', 'STATUS',
       'UNITS', 'Magnitude', 'Subject', 'Group', 'Series_title_1',
       'Series_title_2', 'Series_title_3', 'Series_title_4', 'Series_title_
5'],
      dtype='object')
```

In [14]:
```
df2.drop('Suppressed', axis=1, inplace=True)
```

In [15]:
```
df2.columns
```

Out[15]:
```
Index(['Series_reference', 'Period', 'Data_value', 'STATUS', 'UNITS',
       'Magnitude', 'Subject', 'Group', 'Series_title_1', 'Series_title_2',
       'Series_title_3', 'Series_title_4', 'Series_title_5'],
      dtype='object')
```

In [16]:
```
df2.isnull().sum()
```

Out[16]:

	0
Series_reference	0
Period	0
Data_value	1989
STATUS	0
UNITS	0
Magnitude	0
Subject	0
Group	0
Series_title_1	0
Series_title_2	0
Series_title_3	16174
Series_title_4	20124
Series_title_5	20124

dtype: int64

In [17]:
```
df2.shape
```

Out[17]:
```
(20124, 13)
```

In [18]:
```
df2.drop(['Series_title_3', 'Series_title_4', 'Series_title_5'], axis=1, inp
```

In [19]:
```
df2
```

Out[19]:

	Series_reference	Period	Data_value	STATUS	UNITS	Magnitude	
0	ECTA.S19A1	2001.03	2462.5	F	Dollars	6	Tra (A
1	ECTA.S19A1	2002.03	17177.2	F	Dollars	6	Tra (A
2	ECTA.S19A1	2003.03	22530.5	F	Dollars	6	Tra (A
3	ECTA.S19A1	2004.03	28005.1	F	Dollars	6	Tra (A
4	ECTA.S19A1	2005.03	30629.6	F	Dollars	6	Tra (A
...	...	...	...	...	...	...	
20119	ECTQ.S4AXP	2022.12	32.7	F	Percent	0	Tra (A
20120	ECTQ.S4AXP	2023.03	31.9	F	Percent	0	Tra (A
20121	ECTQ.S4AXP	2023.06	33.0	F	Percent	0	Tra (A
20122	ECTQ.S4AXP	2023.09	33.2	F	Percent	0	Tra (A
20123	ECTQ.S4AXP	2023.12	32.7	F	Percent	0	Tra (A

20124 rows × 10 columns

In [20]:
```
df2.columns
```

Out[20]:
```
Index(['Series_reference', 'Period', 'Data_value', 'STATUS', 'UNITS',
       'Magnitude', 'Subject', 'Group', 'Series_title_1', 'Series_title_
2'],
      dtype='object')
```

In [21]:
```
df2.isnull().sum()
```

Out[21]:

	0
Series_reference	0
Period	0
Data_value	1989
STATUS	0
UNITS	0
Magnitude	0
Subject	0
Group	0
Series_title_1	0
Series_title_2	0

dtype: int64

In [22]:
```
minDataValue = df2['Data_value'].mean()
```

In [23]:
```
df2['Data_value'] = df2['Data_value'].fillna(minDataValue)
```

In [24]:
```
df2.sample(3)
```

Out[24]:

	Series_reference	Period	Data_value	STATUS	UNITS	Magnitude	
5646	ECTM.S19TW	2010.01	2029.5	R	Dollars	6	T
12648	ECTM.S29A1	2016.02	104674806.0	F	Number	0	T
14980	ECTM.S4AXP	2020.06	38.9	F	Percent	0	T

This notebook was converted with convert.ploomber.io