Preprocessing the Socio-Economic Dataset

Data Preprocessing

In [94]:
```
import pandas as pd
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()

df = pd.read_excel('Asansol socio-economic data 1.xlsx')
list1 = list(df.columns)
print(list1)

new_df = df.drop(columns = list1[-3:])
new_df
```

```
['ID', 'Age (Yrs)', 'Gender (Male)', 'Education (<Class 10)', 'Education (Class 10-12)', 'Graduate', 'Higher Edu', 'Income <20k', 'Income 20-40k', 'Income 40-60k', 'Inc
ome 60-80k', 'Income >80k', 'Experience (<1yr)', 'Experience (1-2yr)', 'Experience (2-5yr)', 'Experience (5-10yr)', 'Experience (10-20yr)', 'Experience (20-30yr)', 'Exp
erience (>30yr)', 'Helmet (Full-mask)', 'Unnamed: 20', 'Unnamed: 21', 'Unnamed: 22']
```

Out[94]:

	ID	Age (Yrs)	Gender (Male)	Education (<Class 10)	Education (Class 10-12)	Graduate	Higher Edu	Income <20k	Income 20-40k	Income 40-60k	Income 60-80k	Income >80k	Experience (<1yr)	Experience (1-2yr)	Experience (2-5yr)	Experience (5-10yr)	Experience (10-20yr)	Experience (20-30yr)	Experience (>30yr)	Helmet (Full-mask)
0	1	26	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1
1	2	40	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0
2	3	28	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1
3	4	44	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
4	5	18	1	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
476	477	28	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1
477	478	22	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1
478	479	58	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
479	480	30	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	1
480	481	30	1	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1

481 rows × 20 columns

In [95]:
```
new_df = new_df.drop(columns = ['ID'])
new_df
```

Out[95]:

	Age (Yrs)	Gender (Male)	Education (<Class 10)	Education (Class 10-12)	Graduate	Higher Edu	Income <20k	Income 20-40k	Income 40-60k	Income 60-80k	Income >80k	Experience (<1yr)	Experience (1-2yr)	Experience (2-5yr)	Experience (5-10yr)	Experience (10-20yr)	Experience (20-30yr)	Experience (>30yr)	Helmet (Full-mask)
0	26	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1
1	40	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0
2	28	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1
3	44	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0
4	18	1	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
476	28	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1
477	22	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1
478	58	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
479	30	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	1
480	30	1	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	1

481 rows × 19 columns

In [96]:
```
new_df.describe()
```

Out[96]:

	Age (Yrs)	Gender (Male)	Education (<Class 10)	Education (Class 10-12)	Graduate	Higher Edu	Income <20k	Income 20-40k	Income 40-60k	Income 60-80k	Income >80k	Experience (<1yr)	Experience (1-2yr)	Experience (2-5yr)	Experience (5-10yr)	Experie (10-20
count	481.000000	481.000000	481.000000	481.000000	481.000000	481.000000	481.000000	481.000000	481.000000	481.000000	481.000000	481.000000	481.000000	481.000000	481.000000	481.000
mean	35.877339	0.950104	0.189189	0.305613	0.461538	0.043659	0.241164	0.355509	0.276507	0.108108	0.018711	0.031185	0.081081	0.133056	0.326403	0.303
std	10.662216	0.217957	0.392067	0.461146	0.499038	0.204548	0.428235	0.479166	0.447736	0.310840	0.135644	0.173998	0.273244	0.339989	0.469385	0.460
min	16.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
25%	28.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
50%	35.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
75%	42.000000	1.000000	0.000000	1.000000	1.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000
max	73.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000

In [97]:
```
age = new_df[['Age (Yrs)']]
age

age = scaler.fit_transform(age)
age_df = pd.DataFrame(age,columns = ['Normalised Age (Yrs)'])
age_df.describe()
```

Out[97]:

	Normalised Age (Yrs)
count	4.810000e+02
mean	-1.477220e-16
std	1.001041e+00
min	-1.866219e+00
25%	-7.395780e-01
50%	-8.237053e-02
75%	5.748370e-01
max	3.485327e+00

In [106…
```
Final_df_1 = new_df.drop(columns = ['Age (Yrs)'])

Final_df_1['Normalised Age (Yrs)'] = age_df
list2 = list(Final_df_1.columns)
new_order = [list2[-1]] + list2[0:-1]

Final_df_1 = Final_df_1[new_order]
Final_df_1
```

Out[106]:

	Normalised Age (Yrs)	Gender (Male)	Education (<Class 10)	Education (Class 10-12)	Graduate	Higher Edu	Income <20k	Income 20-40k	Income 40-60k	Income 60-80k	Income >80k	Experience (<1yr)	Experience (1-2yr)	Experience (2-5yr)	Experience (5-10yr)	Experience (10-20yr)	Experience (20-30yr)	Experience (>30yr)	Helm (Ful mas
0	-0.927352	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	
1	0.387063	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	
2	-0.739578	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	
3	0.762611	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	
4	-1.678446	1	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
476	-0.739578	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	
477	-1.302899	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	
478	2.077026	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	
479	-0.551804	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	
480	-0.551804	1	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	

481 rows × 19 columns

In [107…
```
Final_df_1.to_excel('Preprocessed_data_standardscaler.xlsx')
```