Experiment 11

Aim: Data Cleaning and Preprocessing with Pandas: Use the Pandas library in Python to clean and preprocess large datasets, addressing issues such as missing values, outliers, and inconsistent data.

Dataset Preprocesing and Visualization:

Head:

vived	pclass	sex	age	sibsp	parch	fare	embarked
0	3	male	22.0	1	0	7.2500	S
1	1	female	38.0	1	0	71.2833	C
1	3	female	26.0	0	0	7.9250	S
1	1	female	35.0	1	0	53.1000	S
0	3	male	35.0	0	0	8.0500	S

adult_male	deck	embark_town	alive	alone
True	NaN	Southampton	no	False
False	C	Cherbourg	yes	False
False	NaN	Southampton	yes	True
False	C	Southampton	yes	False
True	NaN	Southampton	no	True

Fig 1: Dataset Printing

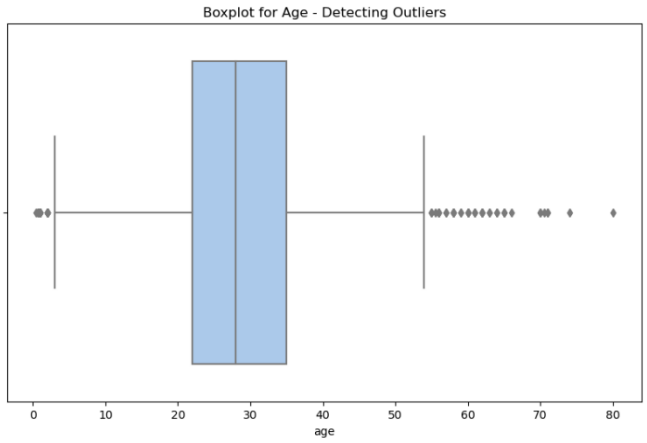

Fig 2: Box Plot for Detecting Outliers

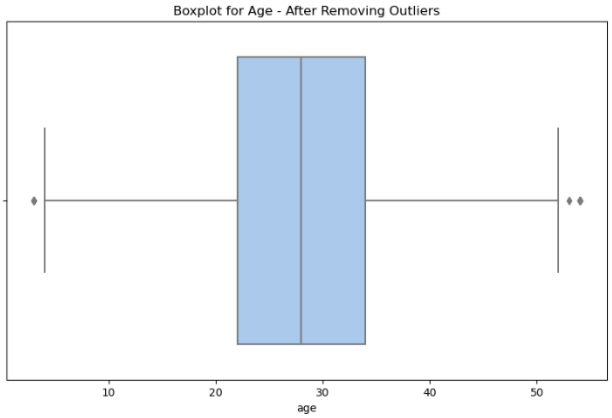

Fig 3: Box Plot after Removing Outliers

Cleaned Dataset Statistics:

	survived	pclass	age	sibsp	parch	fare
count	825.000000	825.000000	825.000000	825.000000	825.000000	825.000000
mean	0.380606	2.341818	28.515152	0.507879	0.357576	31.483615
std	0.485830	0.824096	10.177256	1.090670	0.798599	49.956429
min	0.000000	1.000000	3.000000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	22.000000	0.000000	0.000000	7.895800
50%	0.000000	3.000000	28.000000	0.000000	0.000000	13.416700
75%	1.000000	3.000000	34.000000	1.000000	0.000000	30.070800
max	1.000000	3.000000	54.000000	8.000000	6.000000	512.329200

Fig 4: Descriptive Statistics after Preprocessing

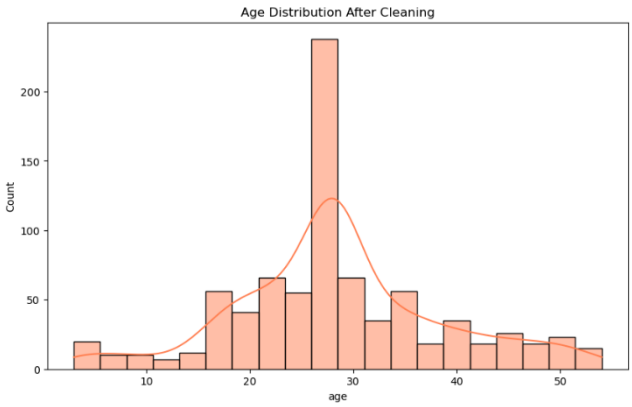

Fig 5: Age Distribution after Cleaning

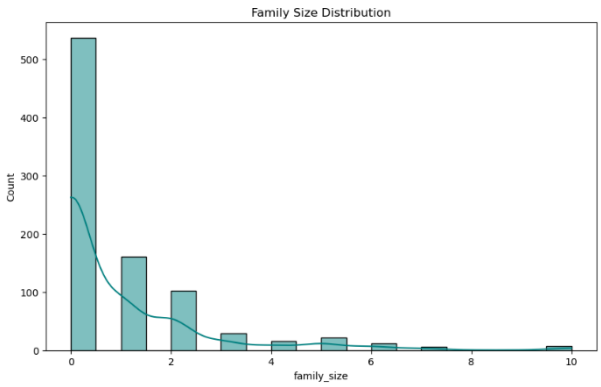

Fig 6: Family Size Distribution