Apply data analysis techniques to extract relevant features from the Titanic dataset for further use in predictive modeling.

Dataset:

Use the Titanic dataset available on Kaggle or through seaborn (sns.load_dataset("titanic")).

Tasks:

Q1. Data Loading & Cleaning

1. Load the Titanic dataset into a Pandas DataFrame.

2. Handle missing values appropriately:

Fill missing age values with the median.

Fill missing embarked with the most frequent value.

Drop the column having too many missing values.

3. Remove any duplicate rows if present.

Q2. Exploratory Data Analysis (EDA)

1. Display summary statistics of numerical and categorical features.

2. Show the correlation matrix of numerical features using a heatmap.

3. Create the following plots:

Survival rate by gender (sex).

Survival rate by passenger class (pclass).

Age distribution of passengers.

Q3. Feature Engineering

1. Convert categorical columns (sex, embarked, class) into numerical using one-hot encoding.

2. Create new derived features:

family_size = sibsp + parch + 1

is_alone = 1 if family size = 1 else 0.

3. Bin the age column into categories (child, teen, adult, senior).

Q4. Dimensionality Reduction

1. Standardize numerical features (age, fare, family_size).

2. Apply Principal Component Analysis (PCA) to reduce them to 2 principal components.

3. Report the explained variance ratio of the components.

Q5. Feature Selection

1. Use SelectKBest (chi2 test) to select the top 5 features most relevant to the survival label (survived).

2. Display the selected features.