EDA Approach

This project is a comprehensive approach to Exploratory Data Analysis (EDA) using a dataset related to heart failure clinical records. Below is a summary of the key steps and insights from the document:

Data Overview

- The dataset contains 5000 rows and 13 columns: age, anemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, and DEATH_EVENT.
- Initial data types and a sample of the first few rows are examined to understand the dataset structure.

Data Cleaning

- Columns 'sex' and 'smoking' are removed as they are deemed less relevant.
- Rows where 'diabetes' equals 0 are filtered out to focus on patients with diabetes.
- Missing values are imputed with mean values for the 'platelets' and 'serum creatinine' columns.

Numerical Data Analysis

- Descriptive statistics, histograms, boxplots, and QQ plots are used to analyze the distribution of numerical variables (age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium).
- Shapiro-Wilk tests indicate that none of these variables follow a normal distribution.

Categorical Data Analysis

- Bar charts and value counts for categorical variables (anaemia, diabetes, high blood pressure, DEATH_EVENT) are analyzed to understand the distribution of these features.
- Key insights include that most patients do not have anaemia, diabetes, or high blood pressure, and the survival rate is higher than the mortality rate in the dataset.

Correlation Analysis

- A correlation heatmap identifies strong positive correlations between serum creatinine and DEATH_EVENT, and time and DEATH_EVENT.
- Negative correlations are found between ejection fraction and DEATH_EVENT, and serum sodium and DEATH_EVENT.

Scatter Plot Analysis

- Scatter plots are generated for various pairs of numerical variables to identify potential relationships.
- Insights include the lack of strong linear relationships between most variable pairs, and that lower ejection fraction and higher serum creatinine levels are more common among patients who experienced a death event.

Principal Component Analysis (PCA)

- PCA is performed to reduce dimensionality and identify principal components that explain the variance in the dataset.
- The first two principal components explain approximately 26.63% of the total variance.

Feature Selection Techniques

- Several feature selection methods are applied, including ANOVA F-test, mutual information, Recursive Feature Elimination (RFE), and Lasso regression.
- The 'time' feature consistently ranks as one of the most important predictors across all methods.

Outlier Detection and Removal

- Outliers are detected and removed using various methods: Z-Score, IQR, Isolation Forest, and DBSCAN.
- The number of outliers detected varies significantly across methods, with DBSCAN detecting the most outliers (1217).

Discretization of Variables

- Continuous variables are discretized into categories using equal-width binning, equal-frequency binning, and custom binning.

- This helps in understanding the distribution of data across different ranges and simplifies the modeling process.

Additional Functions and Techniques

- Functions for data transformation, including changing column types, encoding categorical variables, generating correlation heatmaps, and scatter plots, are implemented to facilitate the EDA process.
- The document includes Python code snippets for each step, demonstrating practical implementation of the described techniques.

Overall, the document provides a detailed and methodical approach to EDA, focusing on data cleaning, numerical and categorical data analysis, correlation analysis, outlier detection, feature selection, and data visualization.