□ Phase 1 Report: Data Preprocessing & EDA

Summary of Work

In this phase, I focused on preparing and exploring the Heart Disease dataset (processed.cleveland.data from UCI). The main goal was to clean the data, handle missing values, and better understand the features before moving to the modeling stages.

1. Dataset Loading

- Loaded the dataset from the UCI repository.
- Assigned meaningful column names for better readability.
- Replaced invalid entries marked as '?' with NaN values.

2. Missing Values Check

The dataset turned out to be almost complete, with only a few missing entries:

- 4 missing values in 'ca' (number of major vessels).
- 2 missing values in 'thal' (thalassemia test result).

3. Exploratory Data Analysis (EDA)

To better understand the data, I created several visualizations:

- Histograms: Showed the distribution of each feature and highlighted skewness in some variables.
- Correlation Heatmap: Helped identify relationships between variables and their correlation with the target.
- Boxplots: Used to detect potential outliers in the numerical features.

4. Preprocessing Pipeline (Scikit-learn)

A reusable preprocessing pipeline was built using Scikit-learn:

- Numeric Features: Missing values were filled with the median, followed by scaling with StandardScaler.
- Categorical Features: Missing values were filled with the most frequent category, and the features were encoded using OneHotEncoder.

5. Final Outputs

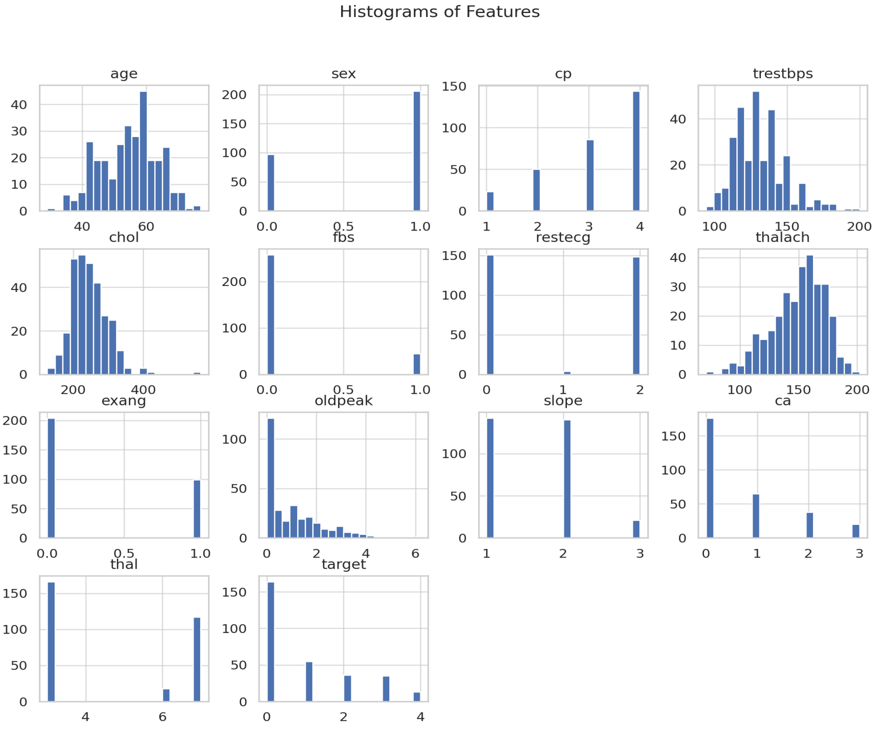
At the end of this phase, I generated two main outputs:

- Cleaned Dataset: heart_disease_cleaned.csv
- Preprocessor Pipeline: preprocessor_pipeline.pkl

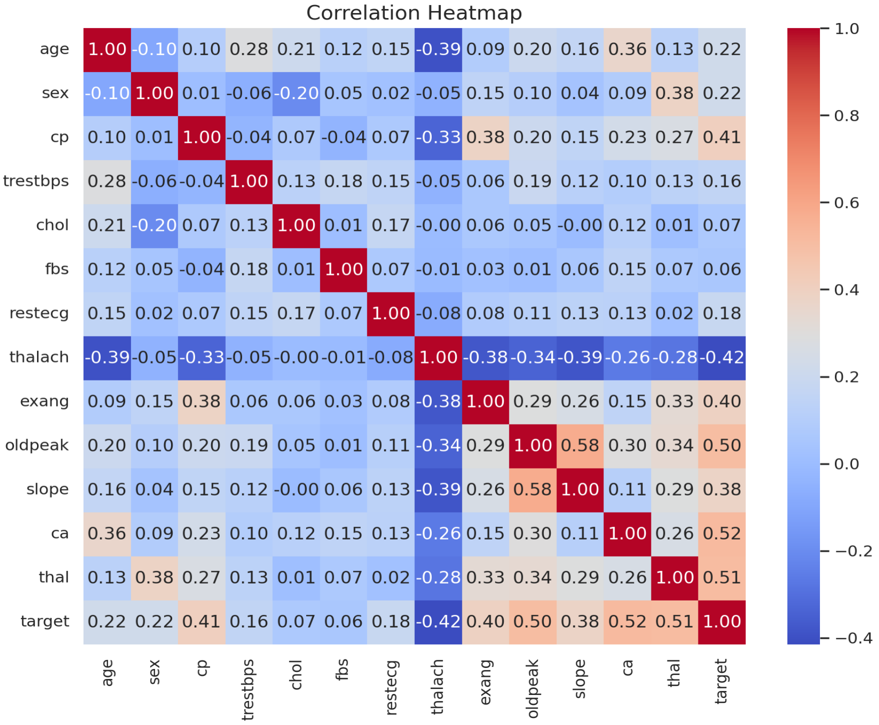
Conclusion

Phase 1 (Data Preprocessing & EDA) is now complete. The dataset is clean, transformed, and ready for Phase 2 (PCA - Dimensionality Reduction).

□ Histograms:

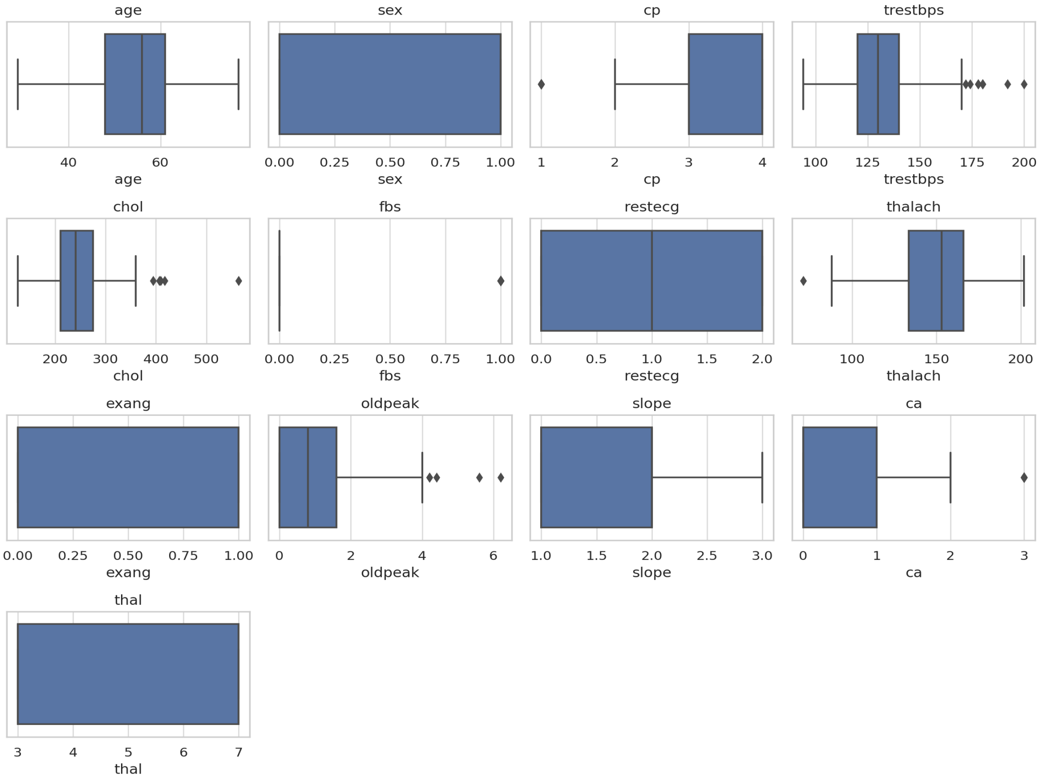


□ Correlation Heatmap:



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.00	-0.10	0.10	0.28	0.21	0.12	0.15	-0.39	0.09	0.20	0.16	0.36	0.13	0.22
sex	-0.10	1.00	0.01	-0.06	-0.20	0.05	0.02	-0.05	0.15	0.10	0.04	0.09	0.38	0.22
cp	0.10	0.01	1.00	-0.04	0.07	-0.04	0.07	-0.33	0.38	0.20	0.15	0.23	0.27	0.41
trestbps	0.28	-0.06	-0.04	1.00	0.13	0.18	0.15	-0.05	0.06	0.19	0.12	0.10	0.13	0.16
chol	0.21	-0.20	0.07	0.13	1.00	0.01	0.17	-0.00	0.06	0.05	-0.00	0.12	0.01	0.07
fbs	0.12	0.05	-0.04	0.18	0.01	1.00	0.07	-0.01	0.03	0.01	0.06	0.15	0.07	0.06
restecg	0.15	0.02	0.07	0.15	0.17	0.07	1.00	-0.08	0.08	0.11	0.13	0.13	0.02	0.18
thalach	-0.39	-0.05	-0.33	-0.05	-0.00	-0.01	-0.08	1.00	-0.38	-0.34	-0.39	-0.26	-0.28	-0.42
exang	0.09	0.15	0.38	0.06	0.06	0.03	0.08	-0.38	1.00	0.29	0.26	0.15	0.33	0.40
oldpeak	0.20	0.10	0.20	0.19	0.05	0.01	0.11	-0.34	0.29	1.00	0.58	0.30	0.34	0.50
slope	0.16	0.04	0.15	0.12	-0.00	0.06	0.13	-0.39	0.26	0.58	1.00	0.11	0.29	0.38
ca	0.36	0.09	0.23	0.10	0.12	0.15	0.13	-0.26	0.15	0.30	0.11	1.00	0.26	0.52
thal	0.13	0.38	0.27	0.13	0.01	0.07	0.02	-0.28	0.33	0.34	0.29	0.26	1.00	0.51
target	0.22	0.22	0.41	0.16	0.07	0.06	0.18	-0.42	0.40	0.50	0.38	0.52	0.51	1.00

□ Boxplots:


age
age
sex
sex
cp
cp
trestbps
trestbps
chol
chol
fbs
fbs
restecg
restecg
thalach
thalach
exang
exang
oldpeak
oldpeak
slope
slope
ca
ca
thal
thal